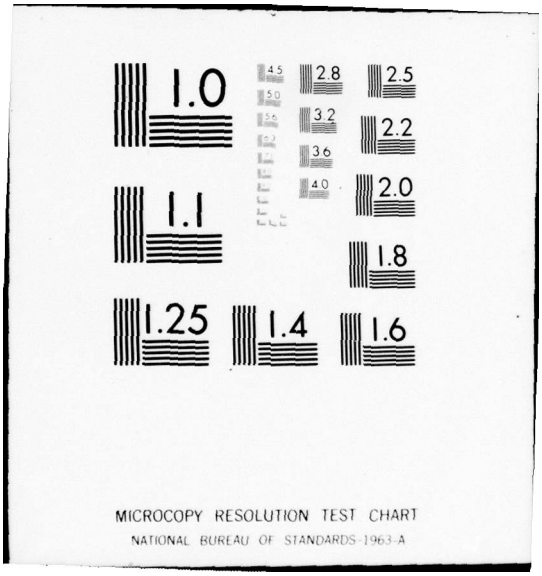
AD-A054 556 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
STUDIES IN THE ANALYSIS OF SERIALLY DEPENDENT DATA.(U)
MAR 78 L C PALLESEN DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1837 NL

1 OF 2
AD
A054556


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(13)



(9) MRC Technical Summary Report, #1837

(6) STUDIES IN THE ANALYSIS OF SERIALLY DEPENDENT DATA.

(10) Lars Chr./Pallesen

(15) DAAG-29-75-C-0024

(11) mar 78

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

(14) MRC-TSR-1837

Mar[illegible]78

(12) 97p.



Received July 8, 1977

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

221 200

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

# STUDIES IN THE ANALYSIS OF SERIALLY DEPENDENT DATA

Lars Chr. Pallesen

Technical Summary Report #1837
March 1978

## ABSTRACT

The analysis of linear models with independent homescedastic, normal noise (white noise) occupies a prominent position in applied statistics. This report is concerned with the linear model analysis of data which cannot be assumed statistically independent because the data have been collected sequentially in time or space.

Assuming that the noise in a linear model (with p-dimensional parameter vector $\underset{\sim}{\beta}$) follows a first order autoregressive scheme (with parameter $\phi$) it is shown in Chapter 2 how in the Bayesian framework inferences can be drawn about $\underset{\sim}{\theta}$ and $\phi$ jointly, conditionally and marginally. Two AR-1 schemes are considered, one covering explosive as well as stationary situations, the other assumes stationarity and reversibility a priori. The important task of choosing an appropriate joint prior distribution for the parameters is given special attention.

Recognizing that the assumption of independence can be a crucial one, it has become a widespread practice in much regression work, where observations are made in time order, to test for serial correlation using the well known Durbin-Watson test. This test corresponds to determining whether a certain estimate $\tilde{\phi}$ of $\phi$ is significantly different from zero. It is shown in Chapter 3 that only in relation to a model lacking a mean do suggested alternative testing procedures show decisively greater empirical power than the DW-test. However it is argued that tests of a null hypothesis of independence should really not be carried out when serial correlation is to be expected. It is demonstrated that inferences about $\underset{\sim}{\theta}$ may be quite misleading if made conditionally on $\phi = 0$ (when that hypothesis is "accepted") or conditionally on $\tilde{\phi}$ (when it is "rejected"). The Bayesian analysis does not suffer from these handicaps. The Bayesian inference about $\underset{\sim}{\theta}$ may be approximated by a conditional inference using the maximum likelihood estimate $\tilde{\phi}$ of $\phi$ ($\tilde{\phi}$ is found by estimating $\phi$ and $\underset{\sim}{\theta}$ simultaneously by least squares), and this conditional analysis is paralleled in sampling theory framework when inferences about $\underset{\sim}{\theta}$ are drawn as if $\phi = \tilde{\phi}$.

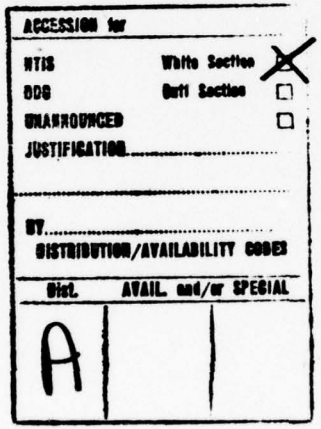
One way in which data analysts have sometimes tried to get around the problem of dependence, is by differencing and then assuming that the errors of the differences are independent. This is equivalent to assuming that $\phi = 1$ in the AR-1 noise model; and the plausibility of this particular value for $\phi$ may be assessed by studying the marginal posterior distribution of $\phi$. For models involving a mean this parameter vanishes for $\phi = 1$. This creates a singularity, and it is shown in Chapter 4, how the density may be found at such (distinct) points. It is also developed in the Bayesian framework, how the appropriateness of differencing may be assessed using posterior model probabilities for two alternative noise models, one assuming the original observations to have AR-1-noise, the other that the differences have AR-1 noise.

AMS(MOS) Subject Classification: Primary 62F15, 62M10
Secondary 62P20, 62N99, 62J05

Key words: Regression, Linear models, Bayesian approach, Serial correlation, Autoregressive noise, Nonstationary noise, Differencing

Work Unit No. 4 - Probability, Statistics and Combinatorics

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 .



## SIGNIFICANCE AND EXPLANATION

One of the most widely used statistical techniques in applied fields like engineering, business, economics, sociology etc. is linear model analysis (analysis of variance, regression analysis, etc.). When the data to be analyzed are collected sequentially in time or space, it turns out, more often than not, that the data at any given point in space or time depend to some extent on the data at neighboring points in space or in instants of time. The data are then said to be serially correlated (autocorrelated). However "standard" methods of analysis usually assume the observations to be statistically independent; if the assumption of independence is incorrect, this will invalidate the analysis and may cause very misleading conclusions.

This report is devoted to studying how serially dependent data may be analyzed, when one extra parameter is incorporated to allow for serial correlation. Specifically it is developed in the Bayesian framework how general linear models may be analyzed when the noise follows a first order autoregressive process (a Markov process). We argue against following the popular practice of checking independence with a Durbin-Watson test whenever serial correlation is feared.

One way to get around the problems of serial correlation which has sometimes appeared helpful is to difference the series, i.e. analyze the increments rather than the original data. The question of whether to difference or not to difference is looked into in a quantitative way.

---

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.


## ACKNOWLEDGEMENTS

Professor G. E. P. Box served as my advisor during the preparation of this thesis, and I deeply appreciate his generous sharing of ideas, humor, wisdom and dedication. I am proud and grateful to count myself among his disciples.

I should also like to express gratitude towards my diverse financial sponsors over the years; in particular I shall acknowledge the vital grants from "Otto Mønsteds Fond" and the never failing backing by "Civilingeniør Frants Allings Legat".

The support in many ways of our close family and our dear friends has been invaluable.

I thank my wife for making my labor worthwhile.

CONTENTS

# CHAPTER 1

## Introduction.

This introductory chapter contains a general discussion of some issues relating to the topics of this thesis, viz: Inference about the linear parameters $\theta_j$, $j = 1,2,\ldots,p$, of a general linear model with first order autoregressive (AR-1) noise, inference about its autoregressive parameter $\phi$, inference about $(\theta_j,\phi)$ jointly, and quantitative assessment of the appropriateness of differencing data series prior to such a linea model analysis.

### 1.1. Models and parametric inference.

There are two requirements a statistical model should satisfy in order to be useful in applications. First it must sufficiently approximate reality; second it should be parsimonious, that is it must be simple enough to lend itself to analysis and interpretation. If this latter requirement is not satisfied, even a realistic model may not be particularly helpful. When such a model is exceedingly complex involving a vast number of parameters, (i) it may be impossible to verify, (ii) it may contain many more unknown parameters than can be estimated with available data, and (iii) its implications may be beyond comprehension anyway.

Of course there are limits to how simple a model can be and still be useful, so it may not always be possible to identify an adequate model among the "standard" model forms, for which satisfactory techniques for analysis have been developed. The purpose of the present work, is to contribute to extending the "stock" of standard model structures, which can be handled satisfactory from a statistical point of view.



The usefulness of a parametric model in applications may be attributed to its ability to extract the relevant information embodied in a data set, and make it accessible for human understanding and use. Typically a useful statistical model incorporates more parameters than those of primary interest. These additional parameters serve in making the model a realistic one, and are sometimes called "nuisance parameters". These nuisance parameters are needed if the inference about the primary parameters is to be valid, but their inclusion has the side effect of complicating the analysis of the model. In the Bayesian framework, (see for example Jeffreys [1961], Box and Tiao [1973]), the difficulties created by nuisance parameters are mainly of a computational nature, since inferences about any subset of parameters can be made from their marginal posterior distribution. This distribution is found from the joint posterior by integrating out those parameters not under consideration. Frequently this integration cannot be carried out analytically, but in principle it can always be done numerically using an electronic computer.

Throughout the present work a Bayesian approach is taken almost exclusively. In this framework the choice of an appropriate prior distribution for the parameters is part of the model specification. The prior may serve as a carrier of already existing information or to incorporate prior judgement in the analysis. However in scientific investigations, one would usually want the prior to be "neutral" or "noninformative" relative to the information in the data. It is not always clear which form the prior should have to express such a desire, and this is among the questions we shall address in this thesis.

## 1.2 Broadening the class of linear models.

Writing a model for a dependent variable $y_i$, i = 1,2,...,n, as

$$\underset{\sim}{y} = \underset{\sim}{\eta} + \underset{\sim}{N} \qquad (1.1)$$

the statistical analysis is particularly simple if the deterministic part of the model $\underset{\sim}{\eta}$ is linear in the parameters, i.e. $\underset{\sim}{\eta} = \underset{\approx}{X}\,\underset{\sim}{\theta}$, where $\underset{\approx}{X}$ is the matrix of independent variables and where $\underset{\sim}{\theta}$ is the p-dimensional vector of linear parameters; and if it is assumed that the stochastic part of the model $\underset{\sim}{N}$ is equal to $\underset{\sim}{\varepsilon}$, where the $\varepsilon_i$'s, i = 1,2,...,n, are i.i.d $N(0,\sigma^2)$ (white noise):

$$\underset{\sim}{y} = \underset{\approx}{X}\,\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}\,. \qquad (1.2)$$

Frequently the assumptions about $\underset{\sim}{y}$ implied by this linear model are found too restrictive to be realistic for data sets met in practice. An adequate representation of the data at hand may often be achieved however, if the class of linear models is broadened by parsimonious use of additional parameters.

To do this we must consider which are most likely ways for the model to be wrong. Box [1976] has suggested that three prime contenders are

(i) need for transformation of the dependent variable,

(ii) allowance for outliers,

(iii) allowance for serial correlation when data are collected sequentially.

The necessary modification in the analysis of these augmentations of the linear model are readily derived using a Bayesian approach.

The idea of transformation of original data is old (see in particular Bartlett [1947] and Tukey [1957]). From the Bayesian viewpoint Box and Cox [1964] analyzed the situation when the model form (1.2) is



appropriate after some suitable power transformation has been applied to the dependent variable. They demonstrated how inference can be made about the transformation and in turn about the linear parameters $\underset{\sim}{\theta}$, by studying the maximized likelihood function as a function of the transformation parameter(s) $\lambda$ ($\underset{\sim}{\lambda}$) or by computing the marginal posterior distribution of $\lambda$ ($\underset{\sim}{\lambda}$).

In some situation the linear model (1.2) should realistically be generalized to allow for the possibility, that one or more of the observations are outliers, i.e. do not conform to the model of the remaining "good" observations. Again the literature on outliers is extensive (in particular see Anscombe [1960], Anscombe and Tukey [1963] and Tukey [1960]). From the Bayesian viewpoint Box and Tiao [1968] assumed in their analysis that a "bad" observation has the same expected value as a good one, but with an inflated variance. Abraham and Box [1975] assumed in an alternative Bayesian analysis of the same problem, that a shift in mean with no change in variance, causes a "bad" value.

It is the purpose of the present work to address some questions relating to the analysis of serially correlated data, i.e. when the augmentation of the linear model should be of the third kind mentioned above.

## 1.3 Analysis of serially correlated data.

Whenever data are collected sequentially in space or in time, serial correlation is to be expected because of the natural continuity of matter and of events developing in time. Thus it is of interest to broaden the class of linear models (1.2) by relaxing the assumption of independence.

At one time there was rather little concern about such dependencies, in particular regression analysis of time series data were commonly conducted under the assumption, that the observations had white noise

errors. Surveys of the literature relating to the effects of serial correlation in regression are given by Anderson [1954] and by Watson [1967]. Realizing that serial correlation could be important Durbin and Watson [1950], [1951] produced a test (the DW-test) which over the years has become widely used. This test is in effect equivalent to estimating a first order autoregressive (AR-1) parameter $\phi$ from the residuals left by an ordinary least squares fit, and seeing whether this esitmate $\hat{\phi}$ is significantly different from zero. Modifications and new tests have bee proposed for example by Abrahamse and Louter [1971], Berenblut and Webb [1973], Durbin [1969], [1970], Grary [1970] and Schmidt [1972], but from the point of view of power, none of these alternatives have been demonstrated to have any decisive advantage over the DW-test, Durbin and Watson [1970], Smith [1976]. Durbin [1970a] argues, that the main incentive for testing a null hypothesis of independence is that the analysis becomes simple if that hypothesis can be "accepted". When it is "rejected", Theil and Nagar [1961] suggest to make inference about the linear parameters $\underset{\sim}{\theta}$ assuming the noise model to be AR-1 with $\phi = \hat{\phi}$ (calculated from the DW statistic).

Cochrane and Orcutt [1949] demonstrated, that ordinary least squares estimates of regression parameters may be very poor if serial correlation is present, and that the situation may be remedied by applying an independence inducing transformation. If the autocorrelation structure is not known, they express scepticism (see also Orcutt and Cochrane [1949]) about the possibility of estimating the structural parameters from ordinary least squares residuals, since they are biased towards randomness; but an iterative procedure for estimating one or two autoregressive parameters is outlined.



Champernowne [1948] suggested a Bayesian approach to estimating the parameters of a $p$-th order autoregressive noise process in a regression model; and Durbin [1960] proposed an asymptotically efficient two step estimation procedure of the regression parameters in a similar model, taking the autoregressive parameters into account. For regression models with AR-1 noise Hildreth [1969] showed, that the maximum likelihood estimates of the parameters are asymptotically efficient and equal to those of Durbin [1960].

Reiser [1975] analyzed the regression model with $p$-th order autoregressive noise applying a structural inference approach, see Fraser [1968]. In this framework inference about (including estimation of) the autoregressive parameters $\underset{\sim}{\phi}$ is made using a marginal likelihood function, and inferences about the regression parameters $\underset{\sim}{\theta}$ are drawn conditional on the estimates of $\underset{\sim}{\phi}$. The structural inference approach to estimation of autoregressive parameters has also been taken by Haq [1970], [1971] and Levenbach [1972].

Zellner and Tiao [1964] showed how marginal inference can be made about regression parameters when the noise follows an AR-1 scheme with unknown parameter $\phi$. Sredni [1970] generalized this Bayesian analysis to cover a $p$-th order autoregressive noise process.

The noise $N_i$ in (1.1) might be considered generated by a member of the very wide class of autoregressive integrated moving average (ARIMA) processes proposed by Box and Jenkins [1970]. Their work suggests that an appropriate ARIMA model of order $(p,d,q)$ should be specified for $N_i$, and then the combined model consisting of both a deterministic (linear) part and a stochastic part should be analyzed. Differencing the data series $d$ times produces a new series whose errors follow an

ARMA($p,q$) model.

Pierce [1971] showed, that estimating the regression parameters and the stochastic parameters of an ARMA($p,q$) noise model simultaneously by least squares asymptotically produces maximum likelihood estimates (assuming Normality). The estimates of the regression parameters are asymptotically uncorrelated with the ARMA parameter estimates, and the latter have the same covariance matrix as would have resulted if the deterministic model component did not exist.

The qualitative use of sample autocorrelation functions to determine when differencing is appropriate, is discussed by Box and Jenkins [1970]. It was pointed out very early by Student [1914], that differencing of sequential data could be employed to eliminate spurious cross correlation between series of observations. Orcutt and Cochrane [1949] suggested, that it might often be helpful in the analysis of economic time series to look at the first differences rather than the original data series. Recently Anderson [1975] studied when the variation of low order ARMA ($p,q$) processes is reduced by differencing.

1.4 Outline of present work.

As far as the present thesis is concerned, attention is limited to situations where the autocorrelated errors may adequately be represented by a one parameter noise model as specified below.

A naturally appealing noise model would be one where the serial correlation fell off smoothly with the distance (or lag) between observations, and a reasonable way for this to happen would seem to be exponentially. Exactly this behavior is obtained from the first order Markov (AR-1) model

$$e_i = \phi e_{i-1} + \epsilon_i \qquad (1.3)$$



when $0 < \phi < 1$ (as before the $\epsilon_i$'s are a white noise series). For this model the theoretical autocorrelations are $\rho_1 = \phi$, $\rho_2 = \phi^2, \ldots, \rho_\ell \phi^\ell, \ldots$.

Adopting the noise model (1.3) we have

$$\underset{\sim}{y} = \underset{\approx}{X}\,\underset{\sim}{\theta} + \underset{\sim}{e} \qquad (1.4)$$

and of course this less restrictive linear model degenerates to the usual independent model (1.2) when the autoregressive parameter $\phi$ is zero. If $\phi$ was known a priori, then the corresponding independence inducing transformation could at once be applied, and inference about the linear parameters $\underset{\sim}{\theta}$ could be made by traditional means, Aitken [1935]; but such prior knowledge would hardly ever exist for real data sets.

Chapter 2 contains a Bayesian analysis of the linear model (1.4). This analysis is a continuation of the work of Zellner and Tiao [1964] and Sredni [1970], and it specifically looks into some difficulties which have arisen about choosing a prior distribution for the parameters. Two AR-1 schemes are considered, one covers explosive as well as stationary situations, the other assumes stationarity and reversibility a priori. It is developed how inferences can be drawn about $\underset{\sim}{\theta}$ and $\phi$ jointly, conditionally and marginally.

Chapter 3 examines alternative ways of probing for serial correlation, viz. the DW testing approach, a likelihood approach and a Bayesian approach. A brief inquiry into the relative power of these measures indicates that in this respect the DW-test is quite good. However our study suggests that the DW testing approach as usually carried out has serious shortcomings of quite a different character. It is demonstrated, (i) that correlation sufficiently large to be serious may not be

detected, and that when this is so inference about $\underset{\sim}{\theta}$ conditional on an "accepted" hypothesis of $\phi = 0$ may give very misleading results. (ii) That, if $\phi = 0$ is "rejected" and inference about $\underset{\sim}{\theta}$ is based on $\phi = \check{\phi}$ (as suggested for example by Thiel and Nagar [1961]) then erroneous conclusions may also be drawn, because $\check{\phi}$ can be a very poor estimate of $\phi$. (iii) That the likelihood and the Bayesian approaches produce not only better estimates for $\phi$, but also approximate confidence or HPD intervals for $\phi$. (iv) That testing should not be carried out under circumstances where serial correlation is to be expected a priori, but inferences should be made from the less restrictive model (1.4). (v) And that the Bayesian inference about the linear parameters $\underset{\sim}{\theta}$ may be approximated by a conditional Bayesian analysis using the maximum likelihood estimate $\hat{\phi}$ of $\phi$; and this conditional analysis is paralleled from a sampling theory point of view if inferences about $\underset{\sim}{\theta}$ are drawn as if $\phi = \hat{\phi}$.

The reason why the DW procedure is inappropriate is further clarified by the Bayesian analysis in Chapter 4 of a particular data set, where the joint posterior distributions of $(\theta_j, \phi)$ $j = 1,2,\ldots,p$ are studied.

One way in which data analysts classically have tried to get around the problems of serial correlation is by analyzing the first differences of the data rather than the original data series themselves; and in some instances this practice has appeared to be helpful. This raises the question of whether to difference or not to difference autocorrelated data series, and this is the main topic of Chapter 4. The AR-1 noise model (1.3) reduces to a random walk model as a special case for $\phi = 1$. So if indeed $\phi = 1$ then the increments $\nabla y_t = y_t - y_{t-1}$



have white noise errors and differencing is the proper course of action. Hence it is of interest to determine how plausible that particular value for $\phi$ appears in light of the data, and this is looked into from a Bayesian point of view. Of course the increments, $\nabla y_t$, may themselves be autocorrelated, and it may be wondered to what extent the data are more consistent with a model having a fixed mean and an AR-1 noise, or with a model lacking a mean and which implies that the first differences have AR-1 noise. A Bayesian assessment of the relative support of these two alternative noise models is also supplied, which may help decide in which way a single parameter for autocorrelation may best be employed.

## CHAPTER 2

### Linear Models with AR-1 Noise: A Bayesian Analysis.

In this chapter we conduct a Bayesian analysis of general linear models, when the noise follows a first order autoregressive scheme. The scope of the analysis is that of parametric inference, primarily about the linear parameters, but also about the autoregressive parameter, although this topic will be treated further in Chapter 3.

This problem was studied earlier by Zellner and Tiao [1964]. They demonstrated how marginal inference about linear regression parameters is possible in the Bayesian framework. Their noise model covering explosive cases as well as stationary ones is nearly equivalent to the one analyzed in the following as far as the likelihood is concerned. However we shall argue, that the prior they employed leads to unacceptable consequences which we have remedied.

Sredni [1970] generalized the noise model to a $p$-th order autoregressive process, but constrained the parameters a priori to cover only stationary situations. On this condition he established the exact likelihood function and an asymptotic noninformative prior for the autoregressive parameters. He did however assume prior independence between the autoregressive parameters on the one hand and the regression parameters on the other hand. Again we believe that this combined prior is inappropriate.

In this chapter the Bayesian treatment of the problem is carried on, with special attention paid to the important task of choosing an appropriate prior distribution. The class of models being considered is a very wide one indeed, covering general linear models, with AR-1 noise which may be stationary or explosive. A subclass (throughout



identified by superscript "*") resting upon an a priori assumption of stationarity is treated separately. Two examples are given for illustration.

### 2.1 The class of models.

We consider the representation of a data vector $\underset{n\times 1}{\underset{\sim}{y}}$ of $n$ observations by the class of models

$$\underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times 1}{\underset{\sim}{\eta}} + \underset{n\times 1}{\underset{\sim}{N}} \tag{2.1}$$

where $\underset{\sim}{\eta}$ is the deterministic part of the model and $\underset{\sim}{N}$ is the stochastic part of the model. Of course $\underset{\sim}{y}$ need not necessarily be the "original" observations, but some transformation of them for example a series of logarithms, or perhaps of first differences.

The structure of $\underset{\sim}{\eta}$ is supposed specified by the general linear form

$$\underset{n\times 1}{\underset{\sim}{\eta}} = \underset{n\times p}{\underset{\approx}{X}}\ \underset{p\times 1}{\underset{\sim}{\theta}} \tag{2.2}$$

where $\underset{\sim}{\theta}' = [\theta_1\theta_2\ldots\theta_p]$ is the vector of $p$ linear parameters, and where

$$\underset{\approx}{X} = [\underset{\sim}{X}_1\underset{\sim}{X}_2\ldots\underset{\sim}{X}_p] = \begin{bmatrix} \underset{\sim}{x}_1' \\ \underset{\sim}{x}_2' \\ \vdots \\ \underset{\sim}{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \cdot & & \cdot \\ \cdot & x_{ij} & \cdot \\ \cdot & & \cdot \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{2.3}$$

is the matrix of the $p$ independent variables which are supposed known exactly. Like the dependent variable, they need not be the "original" variables, but each may be any given function of the "original" ones; and/or they may be indicator variables. Evidently analysis of variance, analysis of covariance, and regression models are all special cases of this general linear model. Furthermore without loss of generality we

can assume that $\underset{n\times p}{X}$ has full column rank $p$.

Concerning the stochastic part of (2.1), we shall consider the case

$$\underset{n\times 1}{N} = \underset{n\times 1}{e} \tag{2.4}$$

where the random errors, $e_i$, follow a first order autoregressive scheme

$$e_i = \phi e_{i-1} + \varepsilon_i \qquad i = 2,3\ldots n \tag{2.5}$$

where it is assumed that the $\varepsilon_i$'s are i.i.d. $N(0,\sigma^2)$.

The first order autoregressive model has mostly been studied over the stationary range $|\phi| < 1$. If stationarity can be assumed a priori, and if further (see for example Anderson [1954]) the error of the first observation is set

$$e_1 = (1-\phi^2)^{-1/2}\varepsilon_1 \tag{2.6}$$

so that

$$\underset{\sim}{\varepsilon} = \underset{\approx}{a}\,\underset{\sim}{e} \tag{2.7}$$

where

$$\underset{n\times n}{a} = \begin{bmatrix} \sqrt{1-\phi^2} & & & 0 \\ -\phi & 1 & & \\ & \ddots & \ddots & \\ 0 & & -\phi & 1 \end{bmatrix} \tag{2.8}$$

then it is seen that $e$ has a multivariate normal distribution $N(0,\ddagger^{\circ})$ with

$$\underset{\approx}{\ddagger}^{\circ} = \sigma^2 \underset{\approx}{a}^{-1}\underset{\approx}{I}\underset{\approx}{a}^{-1'} = \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & & & \\ \phi & 1 & & \text{sym} \\ \vdots & \ddots & \ddots & \\ \phi^{n-1} & \cdots & \phi & 1 \end{bmatrix} = \sigma^2 \underset{\approx}{A} \tag{2.9}$$

where

$$\underset{\approx}{A} = \underset{\approx}{a'}\underset{\approx}{a}\,. \tag{2.10}$$

The doubly symmetric feature of $\ddagger^{\circ}$ (2.9) shows, that (2.6) implies the error series $e_i$, $i = 1,2,\ldots,n$, to be reversible, that is the series taken in reverse order is an identical AR-1 process.



The idea of stationarity, i.e. $\phi$ being between -1 and +1 is really a mathematical concept, but as far as data is concerned there is no particular reason for a cutoff at $\pm 1$; and one should be able to deal with explosive situations where $\phi$ is in fact larger than one. This aspect of the AR-1 model seems however to have been largely ignored.

If stationarity cannot be assumed a priori, instead of (2.5) we write:

$$e_1 = M + \varepsilon_1 \tag{2.11}$$

where $M$ is a starting parameter. In this formulation, the autogressive parameter $\phi$ may take values on the entire real line. If the process is explosive, then $M$ may be viewed as a measure of how far the explosion has progressed when the first observation is taken. If the process is stationary, then $M$ allows for the possibility that the first error, $e_1$, perhaps has an atypical size.

Combining (2.5) and (2.11) we may write

$$\underset{\sim}{\varepsilon} = \underset{\approx}{b}\,\underset{\sim}{e} - \begin{bmatrix} M \\ \underset{\sim}{0} \end{bmatrix} \tag{2.12}$$

where

$$\underset{n\times n}{b} = \begin{bmatrix} 1 & & & 0 \\ -\phi & 1 & & \\ & \ddots & \ddots & \\ 0 & & -\phi & 1 \end{bmatrix} \tag{2.13}$$

and we see, that $\underset{\sim}{e}$ has a multivariate normal distribution with mean vector

$$E(\underset{\sim}{e}) = \underset{\approx}{b}^{-1}\begin{bmatrix} M \\ \underset{\sim}{0} \end{bmatrix} = M\begin{bmatrix} 1 \\ \phi \\ \vdots \\ \phi^{n-1} \end{bmatrix} \tag{2.14}$$

and variance-covariance matrix

$$\underset{\sim}{\Sigma} = \sigma^2 \underset{\sim}{b}^{-1} \underset{\sim}{I} \underset{\sim}{b}^{-1'} = \sigma^2 (\underset{\sim}{b}'\underset{\sim}{b})^{-1} = \sigma^2 \underset{\sim}{B}^{-1}$$

$$= \frac{\sigma^2}{1-\phi^2} \begin{bmatrix} (1-\phi^2) & & & & \text{sym} \\ \phi(1-\phi^2) & (1-\phi^4) & & & \\ \phi^2(1-\phi^2) & \phi(1-\phi^4) & (1-\phi^6) & & \\ \vdots & \vdots & & \ddots & \\ \phi^{n-1}(1-\phi^2) & \phi^{n-2}(1-\phi^4) & \cdots\cdots & & 1-\phi^{2n} \end{bmatrix} \qquad (2.15)$$

We see, that for $\phi > 1$ the unconditional expected value of $e_i$ is growing exponentially (in i), and its variance is also steadily increasing.

In short, the class of models we are concerned with can be written

$$\underset{\sim}{y} = \underset{\sim}{X}\,\underset{\sim}{\theta} + \underset{\sim}{e} \qquad (2.16)$$

with

$$\underset{\sim}{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} M + \varepsilon_1 \\ \phi e_1 + \varepsilon_2 \\ \vdots \\ \phi e_{n-1} + \varepsilon_n \end{bmatrix} \qquad (2.17)$$

where $\underset{\sim}{\varepsilon}$ is sperically normally distributed:

$$p(\underset{\sim}{\varepsilon}) = (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}\,\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}). \qquad (2.18)$$

If stationarity and reversibility can be assumed a priori, then one degree of freedom may be saved by adopting (2.6) instead of (2.11), i.e. by eliminating the parameter M. In terms of (2.17) this elimiation is done by formally replacing M by $((1-\phi^2)^{-1/2}-1)\varepsilon_1$. The resulting subclass of stationary models shall in the following be identified by superscript "o".



## 2.2 Likelihood functions

In this section we derive the likelihood function of the not necessarily stationary model (in the following referred to as the general model), as well as give that of the a priori stationary model. Also the maximum likelihood estimates (MLE) and the information matrices associated with these two likelihood functions are given.

### 2.2.1 Derivation of likelihood functions.

From (2.18) and (2.12) we find

$$p(\underset{\sim}{e}) = (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(\underset{\sim}{e}'\,\underset{\sim}{B}\,\underset{\sim}{e} + M(M-2e_1)) \qquad (2.19)$$

using, that the Jacobian of the transformation (2.12) is unity:

$$\left|\frac{\partial\underset{\sim}{\varepsilon}'}{\partial\underset{\sim}{e}}\right| = \left|\frac{\partial\left[\underset{\sim}{b}\,\underset{\sim}{e} - \begin{bmatrix} M \\ \underset{\sim}{0} \end{bmatrix}\right]'}{\partial\underset{\sim}{e}}\right| = \frac{\partial\,\underset{\sim}{e}'\,\underset{\sim}{b}'}{\partial\underset{\sim}{e}} = |\underset{\sim}{b}'| = 1 \qquad (2.20)$$

Introducing the matrix

$$\underset{(n-1)\times n}{\underset{\sim}{c}} = \begin{bmatrix} -\phi & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -\phi & 1 \end{bmatrix} \qquad (2.21)$$

we may write

$$\underset{\sim}{\varepsilon} = \begin{bmatrix} 0 \\ \underset{\sim}{c}\,\underset{\sim}{e} \end{bmatrix} + \begin{bmatrix} e_1 - M \\ \underset{\sim}{0} \end{bmatrix} \qquad (2.22)$$

or

$$\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = \underset{\sim}{e}'\underset{\sim}{C}\,\underset{\sim}{e} + (e_1-M)^2 \qquad (2.23)$$

where

$$\underset{\sim}{C} = \underset{\sim}{c}'\underset{\sim}{c} \qquad (2.24)$$

so that (2.19) may be expressed as

$$p(\underset{\sim}{e}) = (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(\underset{\sim}{e}'\underset{\sim}{C}\,\underset{\sim}{e} + (e_1-M)^2). \qquad (2.25)$$

Noting that the transformation $\underset{\sim}{e} = \underset{\sim}{y} - \underset{\approx}{X}\,\underset{\sim}{\theta}$ has unit Jacobian we hence have, that the probability density function of $\underset{\sim}{y}$ is

$$p(\underset{\sim}{y}|M,\underset{\sim}{\theta},\phi,\sigma) = (2\pi\sigma^2)^{-n/2}\exp(-\tfrac{1}{2}\sigma^{-2}((\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\theta})'\underset{\approx}{C}(\underset{\sim}{y}-\underset{\sim}{X}\underset{\sim}{\theta})+(y_1-\underset{\sim}{x}_1'\underset{\sim}{\theta}-M)^2)) \tag{2.26}$$

When the observations are considered known, (2.26) is of course also the likelihood function of the general model, and it may be expressed in a particularly simple way in terms of the transformed variables defined by

$$\underset{(n-1)\times 1}{\underset{\sim}{z}} = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix} = \underset{\approx}{C}\,\underset{\sim}{y}\ ; \qquad \underset{(n-1)\times p}{\underset{\approx}{\Xi}} = \underset{\approx}{C}\,\underset{\approx}{X}\ . \tag{2.27}$$

From

$$(\underset{\sim}{y} - \underset{\approx}{X}\,\underset{\sim}{\theta})'\,\underset{\approx}{C}(\underset{\sim}{y} - \underset{\approx}{X}\,\underset{\sim}{\theta}) = (\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta})'(\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta}) \tag{2.28}$$

it is seen that the likelihood (relative to the original observations $\underset{\sim}{y}$) is

$$\ell(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = \sigma^{-n}\exp(-\tfrac{1}{2}\sigma^{-2}((\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta})'(\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta}) + (e_1-M)^2). \tag{2.29}$$

Introducing the further notation

$$SS(\underset{\sim}{\theta},\phi) = (\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta})'(\underset{\sim}{z} - \underset{\approx}{\Xi}\,\underset{\sim}{\theta}) \tag{2.30}$$

the log likelihood of the general model may conveniently be written

$$L(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = -n\log\sigma - \tfrac{1}{2}\sigma^{-2}SS(\underset{\sim}{\theta},\phi) - \tfrac{1}{2}\sigma^{-2}(e_1-M)^2\ . \tag{2.31}$$

For the stationary model it is similarly seen (see for example Sredni [1970]), that the likelihood is

$$\ell^{o}(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = \sigma^{-n}\sqrt{1-\phi^2}\,\exp(-\tfrac{1}{2}\sigma^{-2}(\underset{\sim}{z}^{o} - \underset{\approx}{\Xi}^{o}\underset{\sim}{\theta})'(\underset{\sim}{z} - \underset{\approx}{\Xi}^{o}\underset{\sim}{\theta})) \tag{2.32}$$

where

$$\underset{\sim}{z}^{o} = \begin{bmatrix} z_1 \\ \underset{\sim}{z} \end{bmatrix} = \underset{\approx}{a}\,\underset{\sim}{y}\ ; \quad \underset{\approx}{\Xi}^{o} = \underset{\approx}{a}\,\underset{\approx}{X} \tag{2.33}$$



and using

$$SS^{o}(\underset{\sim}{\theta},\phi) = (\underset{\sim}{z}^{o} - \underset{\approx}{\Xi}^{o}\underset{\sim}{\theta})'(\underset{\sim}{z}^{o} - \underset{\approx}{\Xi}^{o}\underset{\sim}{\theta}) \tag{2.34}$$

the log likelihood may be expressed as

$$L^{o}(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = -n\log\sigma + \tfrac{1}{2}\log(1-\phi^2) - \tfrac{1}{2}\sigma^{-2}SS^{o}(\underset{\sim}{\theta},\phi) \tag{2.35}$$

2.2.2 Maximum likelihood estimates.

Finding the maximum likelihood estimators $(\hat{M},\hat{\underset{\sim}{\theta}},\hat{\phi},\hat{\sigma})$ of the general model, by, as usual, equating the first partial derivatives of $L$ to zero, we find that

$$\frac{\partial L}{\partial M} = 0 \Rightarrow \hat{M} = e_1 = \underset{\sim}{y} - \underset{\sim}{x}_1'\hat{\underset{\sim}{\theta}} \tag{2.36}$$

and subsequently that

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \hat{\sigma}^2 = \frac{SS(\hat{\underset{\sim}{\theta}},\hat{\phi})}{n} \tag{2.37}$$

*hence*

$$L(\hat{M},\hat{\underset{\sim}{\theta}},\hat{\phi},\hat{\sigma}|\underset{\sim}{y}) = -\frac{n}{2}\log\frac{SS(\hat{\underset{\sim}{\theta}},\hat{\phi})}{n} - \frac{n}{2} \tag{2.38}$$

in other words, the problem of maximizing $L$ with respect to $(M,\underset{\sim}{\theta},\phi,\sigma)$, is reduced to that of minimizing $SS(\underset{\sim}{\theta},\phi)$ with respect to $(\underset{\sim}{\theta},\phi)$.

As far as $\underset{\sim}{\theta}$ is concerned this is nothing but an ordinary least squares problem, i.e.

$$\hat{\underset{\sim}{\theta}}(\phi) = (\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\,\underset{\approx}{\Xi}'\underset{\sim}{z} = (\underset{\approx}{X}'\underset{\approx}{C}\,\underset{\approx}{X})^{-1}\,\underset{\approx}{X}\,\underset{\approx}{C}\,\underset{\sim}{y} \tag{2.39}$$

thus

$$SS(\hat{\underset{\sim}{\theta}},\phi) = SS(\hat{\underset{\sim}{\theta}}(\phi),\phi) = \underset{\sim}{z}'\underset{\approx}{\Xi}(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}'\underset{\sim}{z} = \underset{\sim}{y}'(\underset{\approx}{C}-\underset{\approx}{C}\underset{\approx}{X}(\underset{\approx}{X}'\underset{\approx}{C}\underset{\approx}{X})^{-1}\underset{\approx}{X}'\underset{\approx}{C})\underset{\sim}{y} \tag{2.40}$$

and

$$SS(\hat{\underset{\sim}{\theta}},\hat{\phi}) = \min_{\underset{\sim}{\theta},\phi} SS(\underset{\sim}{\theta},\phi) = \min_{\phi} SS(\hat{\underset{\sim}{\theta}},\phi) \tag{2.41}$$

The numerical values of $\hat{\phi}$ and $SS(\hat{\underset{\sim}{\theta}},\hat{\phi})$ may be determined using any of the well known search algorithms for locating the minimum of a "well behaved" function.

Noting that

$$SS(\hat{\underset{\sim}{\theta}},\phi) = (\underset{\sim}{y} - \underset{\approx}{X}\hat{\underset{\sim}{\theta}})' \underset{\approx}{C} (\underset{\sim}{y} - \underset{\approx}{X}\hat{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{e}}' \underset{\approx}{C} \hat{\underset{\sim}{e}} \tag{2.42}$$

so that for given $\hat{\underset{\sim}{\theta}}$

$$\frac{\partial SS(\hat{\underset{\sim}{\theta}},\phi)}{\partial \phi} = 2\phi \sum_{1}^{n-1} \hat{e}_i^2 - 2 \sum_{1}^{n-1} \hat{e}_i \hat{e}_{i+1} \tag{2.43}$$

$$= \hat{\phi} = \frac{\sum_{1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sum_{1}^{n-1} \hat{e}_i^2} \tag{2.44}$$

it is seen, that the iterative procedure depicted in Figure 2.1, (see also Cochrane and Orcutt [1949]) produces the MLE.

For the stationary model it holds that

$$\hat{\sigma}^2 = \frac{SS^o(\hat{\underset{\sim}{\theta}},\phi)}{n} \tag{2.45}$$

$$\hat{\underset{\sim}{\theta}} = (\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^{o})\, \underset{\approx}{\Xi}^{o\prime}\, \underset{\sim}{z}^{o} \tag{2.46}$$

and since

$$L^o(\hat{\underset{\sim}{\theta}},\hat{\phi},\hat{\sigma}|\underset{\sim}{y}) = -\frac{n}{2} \log SS^o(\hat{\underset{\sim}{\theta}},\phi) + \frac{1}{2}\log(1-\hat{\phi}^2) - \frac{n}{2} \tag{2.47}$$

$\hat{\phi}$ may be found from minimizing

$$\left[\frac{SS^o(\hat{\underset{\sim}{\theta}},\phi)}{(1-\phi^2)^{1/n}}\right] = \left[\frac{\underset{\sim}{y}'(\underset{\approx}{A} - \underset{\approx}{A}\,\underset{\approx}{X}(\underset{\approx}{X}'\underset{\approx}{A}\underset{\approx}{X})^{-1}\underset{\approx}{X}'\underset{\approx}{A})\underset{\sim}{y}}{(1-\phi^2)^{1/n}}\right] \tag{2.48}$$

with respect to $\phi$ . If the term $\frac{1}{2}\log(1-\phi^2)$ in the log likelihood function is neglected an approximate expression similar to (2.44) is found:

$$\hat{\phi} \doteq \frac{\sum_{1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sum_{2}^{n-1} \hat{e}_i^2} \tag{2.49}$$



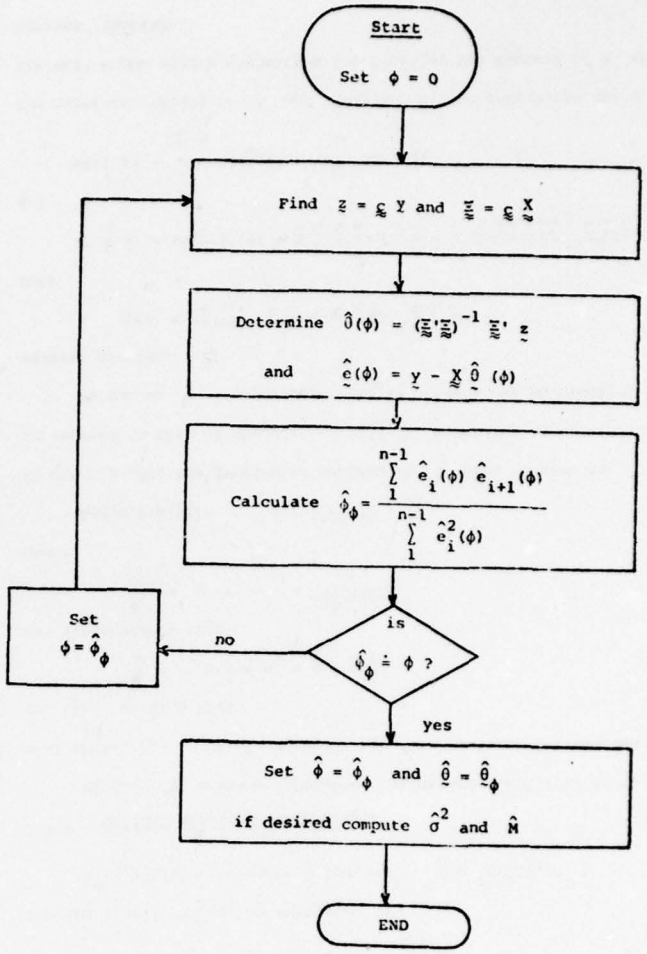



FIGURE 2.1

### 2.2.3 Fisher's information matrix.

The information matrices $\underset{\approx}{\mathcal{J}}$ and $\underset{\approx}{\mathcal{J}}^{o}$ corresponding to the log likelihood functions $L$ and $L^{o}$ are now given

$$\underset{\approx}{\mathcal{J}} = [\mathcal{J}_{ij}] = [-E\frac{\partial^2 L}{\partial\psi_i \partial\psi_j}] \quad \text{with} \quad \psi_i,\psi_j \in \{M,\underset{\sim}{\theta},\phi,\sigma\}$$

$$= \begin{bmatrix} (M) & (\underset{\sim}{\theta}') & (\phi) & (\sigma) \\ \sigma^{-2} & & & \\ \sigma^{-2}\underset{\sim}{x}_1 & \sigma^{-2}(\underset{\approx}{\Xi}'\underset{\sim}{\Xi}+\underset{\sim}{x}_1\underset{\sim}{x}_1') & \text{sym} & \\ 0 & -\sigma^{-2}M\underset{\approx}{X}'\begin{pmatrix}\phi\\ \phi^2-1\\ \vdots\\ \phi^{n-1}-\phi^{n-3}\\ \phi^{n-2}\end{pmatrix} & \mathcal{J}(\phi) & \\ 0 & 0 & 0 & 2n\sigma^{-2} \end{bmatrix} \qquad (2.50)$$

where

$$\mathcal{J}(\phi) = n\left(\frac{1}{1-\phi^2} - \frac{1}{n}\frac{1-\phi^{2n}}{(1-\phi^2)^2} + \frac{1}{n}\frac{M^2}{\sigma^2}\frac{1-\phi^{2(n-1)}}{1-\phi^2}\right) \qquad (2.51)$$

It is seen, that the information about the starting parameter $M$ does not increase with $n$ ; and that the information about $\phi$ not only grows with $n$ , but also is inflated when $|\phi|$ is high and when $M/\sigma$ is large. This is to be expected since under such circumstances the observations will rapidly become very large compared to the error. About $\sigma$ it holds, as usual, that the information is propertional to $n$. As far as $\underset{\sim}{\theta}$ is concerned we observe, that the information matrix essentially involves the *transformed* matrix $\underset{\approx}{\Xi}$ *in the same way as* the original matrix $\underset{\approx}{X}$ appears in the information matrix corresponding to the usual a priori independent linear model.

The information matrix corresponding to the stationary model is



$$\underset{\approx}{\mathcal{J}}^{o} = \begin{bmatrix} (\underset{\sim}{\theta}') & (\phi) & (\sigma) \\ \sigma^{-2}\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^{o} & 0 & 0 \\ 0 & \mathcal{J}^{o}(\phi) & \sigma^{-1}\frac{2\phi}{1-\phi^2} \\ 0 & \sigma^{-1}\frac{2\phi}{1-\phi^2} & 2n\sigma^{-2} \end{bmatrix} \qquad (2.52)$$

where

$$\mathcal{J}^{o}(\phi) = n\left(\frac{1}{1-\phi^2} - \frac{1}{n}\frac{1-3\phi^2}{(1-\phi^2)^2}\right). \qquad (2.53)$$

As far as $\underset{\sim}{\theta}$ and $\sigma$ are concerned this information matrix looks exactly like the one corresponding to the usual a priori independent linear model except the transformed matrix $\underset{\approx}{\Xi}^{o}$ has taken the place of the original matrix $\underset{\approx}{X}$ . The information about $\phi$ is approximately proportional to $n$ and increases also as $|\phi|$ approaches 1.

### 2.3 Prior distributions.

In the Bayesian framework the likelihood function does not alone define a statistical model. It must be complemented with a prior distribution for the parameters; and the two components merge into a complete model.

Being part of the model and not a consequence of it, obviously a prior cannot be logically deduced or "proved", it must be chosen. Still the selection of a prior is arbitrary or subjective only to the extent that model specification must always be. The choice of likelihood function is in principle equally arbitrary or subjective.

Of course not all choices of a model to be tentatively entertained are equally defensible, some make more sense than others.

In this section we consider which priors complement the likelihood functions $L$ and $L^{o}$ in such a way, that the resulting models seem

intuitively reasonable, and in particular do not lead to clearly unacceptable conclusions. The choice of prior can be interpreted as an expression of prior beliefs concerning the parameters. It may also be regarded as a scaling and shaping of the parameter space in the sense, that the prior chosen for the original parameters $\underset{\sim}{\psi}$ may correspond to a locally uniform one in the space of some transformed parameter set $\underset{\sim}{\Psi}$, so that in this parameter space the normalized likelihood function is equal to the posterior distribution. In so preparing the "stage" for the likelihood, it is kept in mind, that it is not critical to "fine tune" remote regions of the parameter space, where it is barely conceivable that any real data would bring the likelihood function .

Only experience can tell whether the suggested priors are conducive to making the overall model useful in practical data analysis.

2.3.1 Prior for the a priori stationary model.

In general the joint prior distribution of $\{\underset{\sim}{\theta},\phi,\sigma\}$ can be factorized as

$$p^{o}(\underset{\sim}{\theta},\phi,\sigma) \propto p^{o}(\underset{\sim}{\theta},\sigma|\phi)\; p^{o}(\phi) \tag{2.54}$$

Focusing attention on the factor $p^{o}(\underset{\sim}{\theta},\sigma|\phi)$ it is recalled, that for given $\phi$ the transformed model expression

$$\underset{\sim}{z}^{o} = \underset{\approx}{\Xi}^{o}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon} \tag{2.55}$$

is nothing but a usual independent linear model involving the parameters $\{\underset{\sim}{\theta},\sigma\}$. Following the argument of Box & Tiao [1973] pp. 25-60, a suitable prior for $\{\underset{\sim}{\theta},\sigma\}$ relative to this model form is

$$p^{o}_{\phi}(\underset{\sim}{\theta},\sigma) \propto \sigma^{-1} . \tag{2.56}$$

The main steps leading to (2.56) are as follows: First an assumption of prior independence between $\underset{\sim}{\theta}$ and $\sigma$ is made. Then treating $\underset{\sim}{\theta}$ and $\sigma$ separately, Jeffreys' rule is invoked. This rule states, that



for a single parameter, $\psi$ , an approximately noninformative prior distribution results from setting

$$p(\psi) \propto (J(\psi))^{1/2} \tag{2.57}$$

where $J(\psi)$ is Fisher's information measure. The rule has the attractive property, that if $\Psi$ is any one to one transformation of $\psi$ , and if $p(\Psi)$ is also determined from (2.57), then

$$p(\psi)d\psi = p(\Psi)d\Psi \tag{2.58}$$

as is easily seen from the relation

$$J(\Psi) = J(\psi)\left(\frac{d\psi}{d\Psi}\right)^{2} . \tag{2.59}$$

Now, if $\Psi$ is a transformation of $\psi$ such that $J(\Psi)$ does not involve $\Psi$ , then the likelihood is approximately data translated, i.e. different outcomes of $\underset{\sim}{y}$ have the effect of changing the location of the likelihood function on the $\Psi$ axis, while leaving its spread approximately unchanged. In this metric a desire to "let the data speak for themselves" may be expressed by choosing a locally uniform prior for $\Psi$ . But if $p(\Psi) \propto k$ is considered a noninformative prior for $\Psi$ , then from (2.58) and (2.59) it follows that a noninformative prior for $\psi$ is

$$p(\psi) = p(\Psi)\left|\frac{d\Psi}{d\psi}\right| \propto (J(\psi))^{1/2} \tag{2.60}$$

The multiparameter version of Jeffereys' rule has the form

$$p(\underset{\sim}{\psi}) \propto \left|\underset{\approx}{J}(\underset{\sim}{\psi})\right|^{1/2} \tag{2.61}$$

If we denote by $\underset{\sim}{\Psi}$ the parameters in the corresponding metric where $p(\underset{\sim}{\Psi})$ is locally uniform, then (2.61) ensures that in this metric the "overall spread" as measured by $\left|\underset{\approx}{J}(\hat{\underset{\sim}{\Psi}})\right|^{-1/2}$ remains approximately constant for different outcomes of $\underset{\sim}{y}$ .

As pointed out by Jeffreys [1961], difficulties can occur in applying the multiparameter rule (2.61). In particular when knowledge is available that certain sets of parameters are independent a priori, this must be

included and the rule applied to the independent sets separately.

In the present multiparameter situation we shall proceed as follows.

Assuming prior independence between $\sigma$ on the one hand and $\underset{\sim}{\theta}$ on the other we have

$$p^o(\underset{\sim}{\theta},\sigma|\phi) = p^o(\underset{\sim}{\theta}|\phi)\, p^o(\sigma|\phi) \tag{2.62}$$

where $p^o(\underset{\sim}{\theta}|\phi)$ is uniform conditionally on $\phi$, i.e. must have the form

$$p^o(\underset{\sim}{\theta}|\phi) \propto g(\phi) \tag{2.63}$$

and on the assumption of prior independence between $\sigma$ and $\phi$

$$p^o(\sigma|\phi) = p^o(\sigma) \propto \sigma^{-1} . \tag{2.64}$$

It may seem arbitrary to parameterize on $\sigma$ which is the standard deviation of the innovations $(\varepsilon_i)$, but in fact if one had parameterized on say $\sigma_e = \sigma/\sqrt{1-\phi^2}$, then precisely the same prior would have resulted:

$$p^o(\sigma) = p^o(\sigma_e)\left|\frac{d\sigma_e}{d\sigma}\right| \propto \left(\frac{\sigma}{\sqrt{1-\phi^2}}\right)^{-1} \frac{1}{\sqrt{1-\phi^2}} = \sigma^{-1} . \tag{2.65}$$

Now as far as $p^o(\underset{\sim}{\theta}|\phi)$ and $p^o(\phi)$ are concerned we find

$$p^o(\underset{\sim}{\theta}|\phi)\, p^o(\phi) = p^o(\underset{\sim}{\theta},\phi) \propto |\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^o|^{1/2}\,(1-\phi)^{-1/2} \tag{2.66}$$

using Jeffreys' rule (2.61) jointly for $\{\underset{\sim}{\theta},\phi\}$, and where the asymptotic approximation

$$p^o(\phi) \overset{\cdot}{\propto} (1-\phi^2)^{-1/2} \tag{2.67}$$

has been adopted. It may be noted, that the nonimformative prior (2.67) for $\phi$, is equivalent to a locally uniform one for $\phi = \text{Arcsin}\,\phi$.

A joint prior distribution for $\{\underset{\sim}{\theta},\phi,\sigma\}$ complementing the likelihood function $\ell^o$ (2.32) is thus produced by multiplying the two separated independent components (2.64) and (2.67):

$$p^o(\underset{\sim}{\theta},\phi,\sigma) \propto |\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^o|^{1/2}\,(1-\phi^2)^{-1/2}\,\sigma^{-1} \tag{2.68}$$



Because multiparameter problems have to be treated with considerable care, it is perhaps valuable to discuss from a somewhat different point of view the prior dependence of $\underset{\sim}{\theta}$ on $\phi$; i.e. the appropriateness of the factor $p^o(\underset{\sim}{\theta}|\phi)$ in the prior.

*If a prior is to be selected which does not prejudice the data, it* might at first sight seem, that this openmindedness could be expressed by a prior implying, that all values of the vector of expected observations $E(\underset{\sim}{y}) = \underset{\sim}{\eta}$ are equally acceptable. This notion implies a locally uniform prior for $\underset{\sim}{\eta}$. However such a choice is inconsistent with the model structure since it ignores the restrictions (2.2) imposed on $\underset{\sim}{\eta}$, as well as it neglects the dependence among the $y_i$'s.

Multiplying $\underset{\sim}{y} = \underset{\sim}{\eta} + \underset{\sim}{e}$ by the independence inducing matrix $\underset{\approx}{a}$ we get

$$\underset{\sim}{z}^o = \underset{\sim}{\zeta}^o + \underset{\sim}{\varepsilon} \tag{2.69}$$

where

$$\underset{\sim}{\zeta}^o = \underset{\approx}{\Xi}^o\underset{\sim}{\theta} \tag{2.70}$$

so that $\underset{\sim}{\zeta}^o$ is restricted to the p-dimensional subspace $W(\phi)$ spanned by the column vectors $\underset{\sim}{\Xi}_1^o, \underset{\sim}{\Xi}_2^o, \ldots, \underset{\sim}{\Xi}_p^o$ of $\underset{\approx}{\Xi}^o$. Hence $\underset{\sim}{\zeta}^o$ is completely described by its projections onto $W(\phi)$. Let $\underset{\approx}{\xi} = |\underset{\sim}{\xi}_1\underset{\sim}{\xi}_2 \cdots \underset{\sim}{\xi}_p|$ be any orthonormal basis for $W(\phi)$. Now consider the vector

$$\underset{\sim}{\Theta} = \underset{\approx}{\xi}'\underset{\sim}{\zeta}^o . \tag{2.71}$$

The j-th element of $\underset{\sim}{\Theta}$ is the length of the projection of $\underset{\sim}{\zeta}^o$ on $\underset{\sim}{\xi}_j$. It is now clear, that a locally uniform prior for $\underset{\sim}{\Theta}$ (conditional on $\phi$) is exactly reflecting the desired openmindedness while being consistent with the explicit restrictions dictated by the model structure. *It follows from*

$$\underset{\sim}{\Theta} = \underset{\approx}{\xi}'\underset{\sim}{\zeta}^o = \underset{\approx}{\xi}'\underset{\approx}{\Xi}^o\underset{\sim}{\theta} \tag{2.72}$$

that a locally uniform prior for $\underset{\sim}{\theta}$ equivalently may be expressed in terms of the following prior for $\underset{\sim}{\theta}$ :

$$p^o(\underset{\sim}{\theta}|\phi) = p^o(\underset{\sim}{\theta}|\phi) \left|\frac{d\underset{\sim}{\theta}}{d\underset{\sim}{\theta}'}\right| \propto |\underset{\approx}{\xi}'\underset{\approx}{\Xi}^o| = |\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\xi}\,\underset{\approx}{\xi}'\underset{\approx}{\Xi}^o|^{1/2} = |\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^o|^{1/2} \tag{2.73}$$

which is exactly the same answer as before.

2.3.2 Prior for the not necessarily stationary model.

The joint prior distribution for $(M,\underset{\sim}{\theta},\phi,\sigma)$ may in general be factorized as

$$p(M,\underset{\sim}{\theta},\phi,\sigma) = p(M,\underset{\sim}{\theta}|\phi,\sigma)\; p(\sigma|\phi)\; p(\phi) \;. \tag{2.74}$$

It seems appropriate to assume independence a priori between $M$ and $\phi$; also we shall assume as before that $\sigma$ is independent a priori of the remaining parameters. On this basis we may write

$$p(M,\underset{\sim}{\theta},\phi,\sigma) = p(M)\; p(\underset{\sim}{\theta}|M,\phi)\; p(\phi)\; p(\sigma) \tag{2.75}$$

Following the same line of argument as was put forth in the previous subsection we get

$$p(\sigma) \propto \sigma^{-1} \tag{2.76}$$

and

$$p(\underset{\sim}{\theta},M|\phi) = p(\underset{\sim}{\theta}|M,\phi)\; p(M) \propto |\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2} \tag{2.77}$$

where clearly the prior for the linear starting parameter $M$ is set locally uniform.

Thus a prior distribution for $(M,\underset{\sim}{\theta},\phi,\sigma)$ complementing the likelihood function $\ell$ (2.29) has the form

$$p(M,\underset{\sim}{\theta},\phi,\sigma) \propto p(\phi)|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}\,\sigma^{-1} \tag{2.78}$$

where the question of choosing an appropriate prior for $\phi$ has been left open. This question shall be returned to in Section 2.5.



## 2.4 Posterior distributions.

At this point Bayes theorem is invoked to establish the posterior density functions of the parameters in the two models, as the normalized product of their respective likelihoods and their complementing priors.

Specifically we have for the general model

$$\begin{aligned} p(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) &\propto p(M,\underset{\sim}{\theta},\phi,\sigma)\,\ell(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) \\ &\propto p(\phi)|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}\,\sigma^{-(n+1)}\exp(-\tfrac{1}{2}\sigma^{-2}(SS(\underset{\sim}{\theta},\phi) + (\varepsilon_1 - M)^2)) \end{aligned} \tag{2.79}$$

and for the stationary model

$$\begin{aligned} p^o(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) &\propto p^o(\underset{\sim}{\theta},\phi,\sigma)\,\ell^o(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) \\ &\propto |\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^o|^{1/2}\,\sigma^{-(n+1)}\exp(-\tfrac{1}{2}\sigma^{-2}SS^o(\underset{\sim}{\theta},\sigma)) \;. \end{aligned} \tag{2.80}$$

From a Bayesian point of view, all information which the data $\underset{\sim}{y}$ may possess about the parameters $(M,\underset{\sim}{\theta},\phi,\sigma)$ (or alternatively $(\underset{\sim}{\theta},\phi,\sigma)$) is explicitly expressed as the posterior distribution $p(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y})$ (or $p^o(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y})$). And all parametric inference consists of interpreting that distribution.

Usually interest is focused on a subset of parameters in which case the joint distribution is an unnecessary complicated carrier of information. Of course using the Bayesian approach marginal inference is made from considering the relevant marginal densities, which in principle are easily found by integrating out the nuisance parameters.

In the context of the present models the parameters $M$ and $\sigma$ will typically be of little if any interest by themselves. Accordingly we shall proceed by eliminating $M$ and $\sigma$ to establish the marginal distributions of $\underset{\sim}{\theta}$ and $\phi$ jointly. In following steps involving further integration, we shall see how inferences are made about $\underset{\sim}{\theta}$, or a subset of $\underset{\sim}{\theta}$, and about $\phi$ individually.

2.4.1 Marginal posterior distribution of $\{\underset{\sim}{\theta},\phi\}$ jointly.

Recognizing, that $p(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y})$ (2.79) is a Normal function with respect to $M$, this parameter may be integrated out to yield

$$p(\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) \propto p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}\sigma^{-n}\exp(-\tfrac{1}{2}\sigma^{-2}SS(\underset{\sim}{\theta},\phi))\int_R \sigma^{-1}\exp(-\tfrac{1}{2}\sigma^{-2}(M-e_1)^2)dM$$

$$\propto p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}\sigma^{-n}\exp(-\tfrac{1}{2}\sigma^{-2}SS(\underset{\sim}{\theta},\phi)) \qquad (2.81)$$

This density is of the Gamma form with respect to $\sigma$, so integrating over that parameter gives

$$p(\underset{\sim}{\theta},\phi|\underset{\sim}{y}) \propto p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}\int_0^\infty \sigma^{-n}\exp(-\tfrac{1}{2}\sigma^{-2}SS(\underset{\sim}{\theta},\phi))d\sigma$$

$$\propto p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}(SS(\underset{\sim}{\theta},\phi))^{-\frac{n-1}{2}} \qquad (2.82)$$

Turning to the stationary model we find similarly:

$$p^o(\underset{\sim}{\theta},\phi|\underset{\sim}{y}) \propto \left|\underset{\approx}{\Xi}^o{}'\underset{\approx}{\Xi}^o\right|^{1/2}(SS^o(\underset{\sim}{\theta},\phi))^{-n/2} \qquad (2.83)$$

2.4.2 Marginal posterior distribution of $\phi$.

Integrating over $\underset{\sim}{\theta}$ we find

$$p(\phi|\underset{\sim}{y}) \propto \int_{R^p} p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}(SS(\underset{\sim}{\theta},\phi))^{-\frac{n-1}{2}}d\underset{\sim}{\theta}$$

$$= p(\phi)\left|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}\right|^{1/2}\int_{R^p}(SS(\hat{\underset{\sim}{\theta}},\phi) + (\underset{\sim}{\theta}-\hat{\underset{\sim}{\theta}})\underset{\approx}{\Xi}'\underset{\approx}{\Xi}(\underset{\sim}{\theta}-\hat{\underset{\sim}{\theta}}))^{-\frac{n-1}{2}}d\underset{\sim}{\theta} \qquad (2.84)$$

Excluding the point(s) $\phi^*$ (if any) which makes $\underset{\approx}{\Xi}'\underset{\approx}{\Xi}$ singular, this t-form expression integrates to

$$p(\phi|\underset{\sim}{y}) \propto p(\phi)\,(SS(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p-1}{2}} \qquad (2.85)$$

with $-\infty < \phi < \infty$.

This exclusion is a matter of formality only since the probability of $\phi$ being equal to the distinct point(s) $\phi^*$ is zero (see Section 3.2 of Chapter 3 and Section 4.2 of Chapter 4).



Similarly for the stationary model we find

$$p^o(\phi|\underset{\sim}{y}) \propto (SS^o(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p}{2}} \qquad (2.86)$$

with $-1 < \phi < 1$

Attention is at this point directed to Appendix B, where it is studied what the consequences would have been for $p(\phi|\underset{\sim}{y})$ had prior independence been assumed between $\underset{\sim}{\theta}$ and $\phi$.

2.4.3 Marginal posterior distribution of $\underset{\sim}{\theta}$.

The marginal posterior distribution of $\underset{\sim}{\theta}$ relative to the general model is found as

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int_R p(\underset{\sim}{\theta},\phi|\underset{\sim}{y})d\phi \qquad (2.87)$$

but here it does not seem possible to perform the integration analytically, so an explicit expression of $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ in terms of well known functions of $\underset{\sim}{y}$ is unavailable.

Utilizing the factorization

$$p(\underset{\sim}{\theta},\phi|\underset{\sim}{y}) = p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y})\;p(\phi|\underset{\sim}{y}) \qquad (2.88)$$

the integration (2.87) may be carried out conveniently by numerical methods on an electronic computer.

The second factor in (2.88) shall be looked into in Section 2.5 and the first factor is the posterior distribution of $\underset{\sim}{\theta}$ given $\phi$, which from (2.84) is seen to be a multivariate t-distribution, specifically

$$p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y}) \sim t_p(\hat{\underset{\sim}{\theta}},s^2(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1},\nu)$$

with

$$\hat{\underset{\sim}{\theta}} = (\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}'\underset{\sim}{z} \qquad (2.89)$$

$$\nu s^2 = SS(\hat{\underset{\sim}{\theta}},\phi)$$

and

$$\nu = n-p-1\ .$$

Hence substituting (2.88) into (2.87) we see, that $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ may be regarded as a weighted sum of t-distributions, where $p(\phi|\underset{\sim}{y})$ acts as a weight function. In this interpretation the special case of independence is equivalent to setting $p(\phi|\underset{\sim}{y})$ equal to an impulse function, $I_{(\phi=0)}$.

Of course a numerical integration consists of summing the integrand in small increments of $\phi$, multiplied by the increment $\Delta\phi$, i.e. setting

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) \doteq \Delta\phi \sum_{\Delta\phi} p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y})\ p(\phi|\underset{\sim}{y}). \tag{2.90}$$

Since $p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y})$ is known exactly including the normalizing constant, and interpreting $p(\phi|\underset{\sim}{y})$ as a weight function, we need not find the normalizing constant of that distribution very precisely in order to assure that $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ integrates to one. The only requirement is, that the weights, being proportional to $p(\phi|\underset{\sim}{y})$, add up to one in the actual summation (2.90). In practice $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ may often be approximated extremely well by a weighted sum of say 5 t-distributions, corresponding to 5 equidistant $\phi$-values covering the interval where the body of $p(\phi|\underset{\sim}{y})$ is located.

Frequently it is desired to make inference about a subset of $\underset{\sim}{\theta}$. Suppose for example, that the r-dimensional subvector $\underset{\sim}{\theta}_1$ is of separate interest, where without loss of generality we may assume

$$\underset{\sim}{\theta} = \begin{pmatrix} \underset{\sim}{\theta}_1 \\ \underset{\sim}{\theta}_2 \end{pmatrix} \tag{2.91}$$

then we find that



$$\begin{aligned} p(\underset{\sim}{\theta}_1|\underset{\sim}{y}) &= \int_{R^{p-r}} p(\underset{\sim}{\theta}|\underset{\sim}{y})\,d\underset{\sim}{\theta}_2 \\ &= \int_R p(\phi|\underset{\sim}{y}) \int_{R^{p-r}} p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y})\,d\underset{\sim}{\theta}_2\,d\phi \\ &= \int_R p(\underset{\sim}{\theta}_1|\phi,\underset{\sim}{y})\ p(\phi|\underset{\sim}{y})\,d\phi \end{aligned} \tag{2.92}$$

where as before $p(\underset{\sim}{\theta}_1|\phi,\underset{\sim}{y})$ is a t distribution, specifically

$$p(\underset{\sim}{\theta}_1|\phi,\underset{\sim}{y}) \sim t_r(\hat{\underset{\sim}{\theta}}_1, s^2\,\underset{\approx}{D}_{11}, \nu) \tag{2.93}$$

with $\hat{\underset{\sim}{\theta}}_1$, $s^2$ and $\nu$ as before, equation (2.89), and where $\underset{\approx}{D}_{11}$ comes from

$$(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1} = \begin{matrix} \\ p \\ p-r \end{matrix} \begin{matrix} \begin{matrix} r & p-r \end{matrix} \\ \begin{bmatrix} \underset{\approx}{D}_{11} & \underset{\approx}{D}_{12} \\ \underset{\approx}{D}_{12} & \underset{\approx}{D}_{22} \end{bmatrix} \end{matrix}. \tag{2.94}$$

So again $p(\underset{\sim}{\theta}_1|y)$ is a weighted sum of t-distributions conditional on $\phi$, the weights being proportional to the relative credibility of the $\phi$-values as expressed by the marginal posterior distribution of $\phi$.

Exactly parallel results hold obviously for the stationary model. The corresponding formulae relative to this model are generated by substituting n for n-1 and inserting superscript "$^0$" wherever applicable.

## 2.5 Marginal prior and posterior for $\phi$ in the not necessarily stationary model.

Turning to the question of choosing a prior for $\phi$, two possibilities are considered. One is an approximately noninformative prior $p_n(\phi)$ of the Jeffreys type, the other is a locally uniform one

$$p_u(\phi) \propto k. \tag{2.95}$$

How informative this prior is may be judged by comparison to $p_n(\phi)$, and we shall argue below, that $p_u(\phi)$ is not so unreasonable a choice as it

might first seem. If we adopted this prior (2.95), we get from (2.85):

$$p_u(\phi|\underset{\sim}{y}) \propto (SS(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p-1}{2}} \tag{2.96}$$

which incidentally looks much like $p^0(\phi|\underset{\sim}{y})$ (2.86) of the stationary model.

It turns out, that $p_u(\phi|\underset{\sim}{y})$ can be approximated closely by a t-distribution with n-p-2 degrees of freedom; which is equivalent to saying that in the vicinity of $\phi = \hat{\phi}$, $SS(\hat{\underset{\sim}{\theta}},\phi)$ is very nearly a quadratic function.

According to Jeffreys' rule (2.60), $p_n(\phi)$ should locally be proportional to $J(\phi)^{1/2}$, where (from the shorthand formula (2.51)):

$$J(\phi) = (n-1)+(n-2)\phi^2 + (n-3)\phi^4+\ldots+ 2\phi^{2(n-3)}+\phi^{2(n-2)} + \frac{M^2}{\sigma^2}(1+\phi^2 + \ldots + \phi^{2(n-2)}) \tag{2.97}$$

with $-\infty < \phi < \infty$.

If $|\phi|$ is small ($|\phi| < 1$), then $J(\phi)$ is dominated by the low order terms. In this case we see, that unless the quantity $M^2/(n\sigma^2)$ is approaching unity, $J(\phi)$ may be approximated well by

$$J(\phi) \doteq (n-1)+(n-2)\phi^2 + (n-3)\phi^4 + \ldots + \phi^{2(n-2)} \quad . \tag{2.98}$$

In practice $M^2/\sigma^2$ should not often be very large for a truly stationary process. And if that contingency is feared in an experimental situation, any reasonable investigator would intuitively counter by making a comparatively larger number of observations ("allowing the system to settle").

So on this assumption a locally approximately noninformative prior looks like

$$p_n(\phi) \propto ((n-1)+(n-2)\phi^2+\ldots+\phi^{2(n-2)})^{1/2} \quad . \tag{2.99}$$



If on the other hand the process is explosive, then the high order terms dominate $J(\phi)$, (2.97). Of course if $M^2/\sigma^2$ is small (i.e. the explosion has been monitored from its beginning) then (2.98) is still valid. But even if $M^2/\sigma^2$ is large, $p_n(\phi)$ of (2.99) would seem a reasonable approximation, since now a noninformative prior should look like $(1+\phi^2+\ldots+\phi^{2(n-2)})^{1/2}$ or $(\phi^{2(n-2)})^{1/2}$, which is not much different from (2.99) for large $\phi$.

So in any event we shall adopt (2.99) as a locally approximately noninformative prior for $\phi$. Figure 2.2 shows $p_n(\phi)$ plotted for n = 6, 15, 25 and 60; the four curves have been scaled to go through 1 for $\phi = \pm 1$. (The $q(\phi)$ curves are explained below).

Unfortunately a direct multiplication by this noninformative prior in (2.85) does not yield meaningful results. It turns out (see the example in Section 2.6.1) that if it is attempted to derive $p(\phi|\underset{\sim}{y})$ as $p_n(\phi)(SS(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p-1}{2}}$, then this product blows up for large $\phi$ values (positive or negative), where the likelihood function is approaching zero. The cause of this calamity may be explained by Figures 2.3 and 2.4:

First it is reminded, that the noninformative prior for $\phi$ is equivalent to a locally uniform prior for some $\Phi = \Phi(\phi)$, where the $\Phi$-metric is such that the likelihood function is approximately data translated. In a single parameter situation the effect of a noninformative prior may be illustrated by drawing likelihood curves in the metric induced by that prior. Of course this is not possible to do for multiparameter likelihood functions. What can be done, is drawing marginal posterior distributions for $\phi$ in the original as well as the data

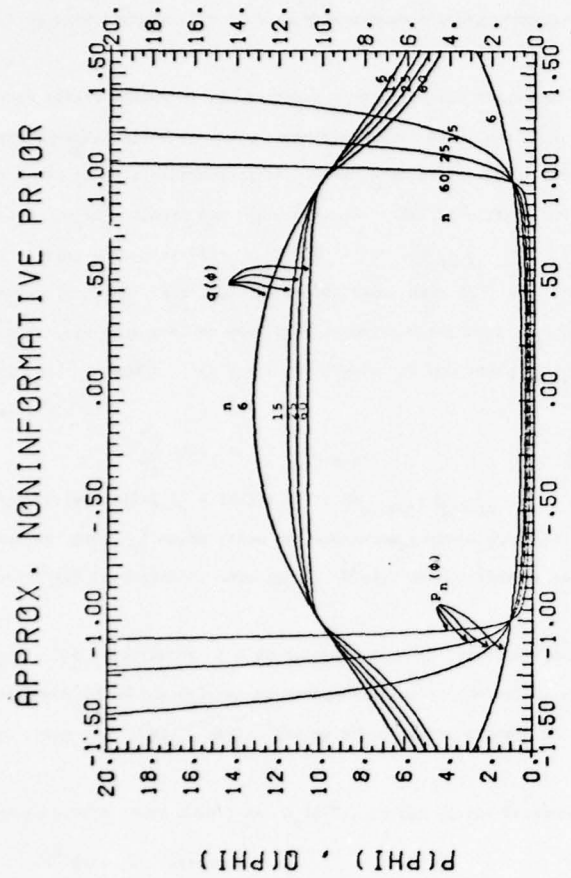



FIGURE 2.2



translating metric; this is done in Figures 2.3 and 2.4 for sample sizes n = 15 and 25 respectively. Actually an alternative argument for producing a noninformative prior, may consist of determining $\Phi(\phi)$ so that the (expected) spread of the posterior distribution of $\Phi$ does not depend on $\phi$. This approach is sketched in Appendix A, and as it turns out the very same prior for $\phi$ is derived that way.

The construction of Figures 2.3 and 2.4 is explained in detail in Appendix A, let it suffice here to point out, that the five solid density curves, $p_u(\phi|\underset{\sim}{y})$, in 2.3a and 2.4a come from (2.96) i.e. correspond to a uniform prior for $\phi$; and they are the results of "data" generated with the true value of $\phi$ being equal to -1.2, -.6, .0, .6 and 1.2. Clearly the extreme curves ($\phi = \pm 1.2$) are much more concentrated than the ones in between, signifying that in the original metric a change in $\hat{\phi}$ (the mode of these curves) not only shifts the posterior along the $\phi$-axis but also changes its shape substantially. In contrast Figures 2.2b and 2.4b show that in the $\Phi$-metric induced by (2.99) the posterior of $\Phi$ remains comparatively unchanged in spread for various values of $\hat{\Phi} = \Phi(\hat{\phi})$. Because the spread of the posterior distribution (or curvature of the likelihood function) not only depends on $\hat{\phi}$ but more generally on the data, sampling variation in the spread of $p(\phi|\underset{\sim}{y})$ is to be expected. The adequacy of the $\Phi$-metric is supported by the lack of pattern from plot to plot of $p(\Phi|\underset{\sim}{y})$ (see also plots in Appendix A); so apparently the prior $p_n(\phi)$ is doing what it is supposed to do. But looking more closely at Figures 2.3b and 2.4b it is observed, that the posteriors $p(\Phi|\underset{\sim}{y})$ in the data translating metric, unlike $p_u(\phi|\underset{\sim}{y})$ in Figures 2.3a and 2.4a are not nearly symmetric, but possess heavy tails in one direction or the other. The principle of noninformative priors

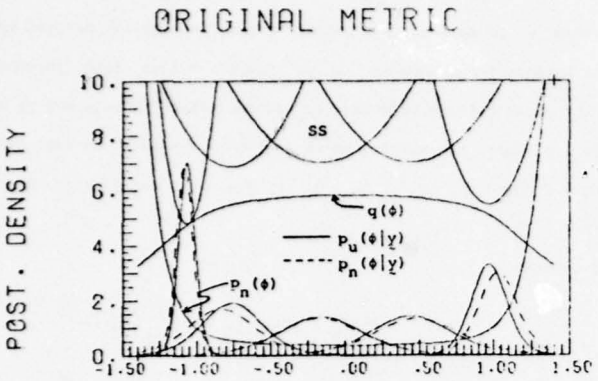



FIGURE 2.3a

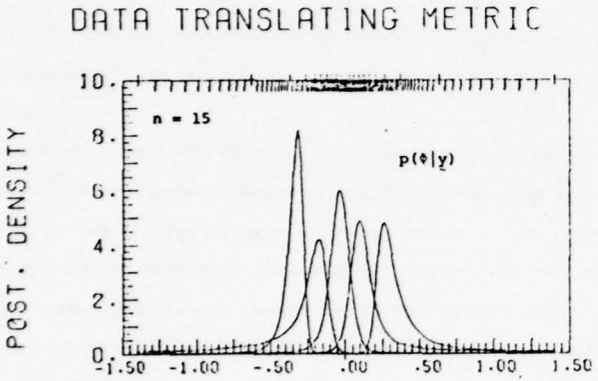



FIGURE 2.3b



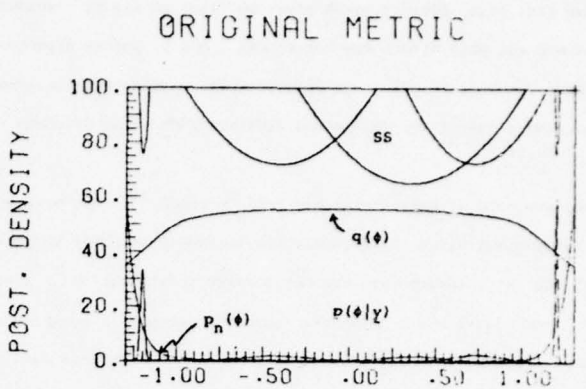



FIGURE 2.4a

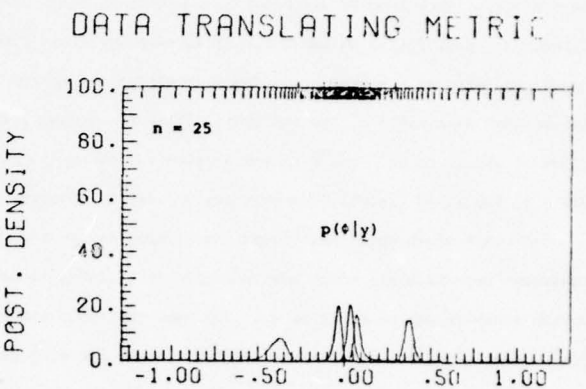



FIGURE 2.4b

does not go beyond the second moment however. Strictly speaking the argument is a local one applicable only in the vicinity of $\hat{\phi}$, where, based on the expected value of a second derivative at this point (used as a measure of spread) the locally approximately noninformative prior has the effect relative to a locally uniform one, of increasing the spread somewhat accompanied by a moderate shift in the mode away from the origin. Hence while a locally uniform prior for $\phi$ may still be advocated, it is apparently necessary to ensure in some way, that the uniform prior is only applied near $\hat{\phi}$ and does not extend to remote tail areas.

Specifically it is suggested, that the modifying effect of $p_n(\phi)$ on $p_u(\phi|\underset{\sim}{y})$ may be achieved within the t-form the following way. It is noted, that multiplying $SS(\hat{\underset{\sim}{\theta}},\phi)^{-\frac{n-p-1}{2}}$ by $p_n(\phi)$ is equivalent to multiplying $SS(\hat{\underset{\sim}{\theta}},\phi)$ by $q(\phi)$ in (2.85), where

$$q(\phi) = p_n(\phi)^{-\frac{2}{n-p-1}} = \begin{cases} 10\left(\frac{2}{n-1}\,\frac{1}{1-\phi^2}\left(1-\frac{1}{n}\,\frac{1-\phi^{2n}}{1-\phi^2}\right)\right)^{-\frac{1}{n-p-1}} & \text{when } \phi \neq \pm 1 \\ 10 & \text{when } \phi = \pm 1 \end{cases} \tag{2.100}$$

$q(\phi)$ is plotted in Figure 2.2 for $n = 6, 15, 25, 60$ and $p = 0$; and scaled like the expression (2.100) to go through 10. for $\phi = \pm 1$. Thus if the product $q(\phi)\,SS(\hat{\underset{\sim}{\theta}},\phi)$ is approximated by a second order polynomial $Q(\phi)$ in the vicinity of $\hat{\phi}$, we have the following t-approximation for the marginal posterior distribution of $\phi$ based locally on the noninformative prior $p_n(\phi)$[†]:

[†] The mechanics of this suggested procedures are spelled out in Section 3.4 of Chapter 3.



$$p_n(\phi|\underset{\sim}{y}) \overset{\cdot}{\propto} (Q(\phi))^{-\frac{n-p-1}{2}} . \tag{2.101}$$

The posterior distributions (2.101) are drawn as the broken curves in Figures 2.3a and 2.4a. (In the latter figure $p_n(\phi|\underset{\sim}{y})$ and $p_u(\phi|\underset{\sim}{y})$ are almost coincidal even for $\phi = \pm 1.2$ where the dramatic upturn in the prior, Figure 2.2, might perhaps leave the erroneous impression that the effect of the locally noninformative prior grows with $n$).

Although $p_n(\phi)$ is the advocated (local) prior for $\phi$, arguments exist in favor of the uniform prior $p_u(\phi)$. It is simple in application, not only because it is identical for all $n$, but more importantly because direct multiplication does not create any problem, and therefore an exact posterior density function can be established. Further it may be questioned, whether a noninformative prior really reflects a realistic prior attitude about $\phi$. Formally $\phi$ is defined over the entire interval $(-\infty,\infty)$, but in real life $\phi$ will hardly ever exceed unity by much. For example a value of $\phi = 3$ is rather unthinkable. Such a process would be so violently explosive, that even for a fairly small sample the smallest (i.e. the first) observation and the largest (i.e. the last) observation could be orders of magnitude apart, while the variance of the disturbance $\varepsilon_i$ added to each new observation is constant (see (2.17)).

Compared to an approximately noninformative prior, a uniform prior is somewhat conservative concentrating the prior probability density more closely around $\phi = 0$. The larger the sample size the greater the difference. It may be felt, at least qualitatively, that this departure reflects a reasonable decreasing willingness to accept larger values of $\phi$ a priori, as $n$ increases.

Finally it is noted, that $p_n(\phi)$ and $p_u(\phi)$ are much alike for small $n$, to the point of being identical for the smallest possible sample, $n = 2$. And as $n$ grows large, the likelihood becomes increasingly dominating in the posterior anyway. Hence even if $p_n(\phi)$ is recognized as the superior choice of prior, from the point of view of being unprejudiced relative to the information supplied by the data, the use of $p_u(\phi)$ may still be contemplated, as the practical difference between the two can be expected to be small.

## 2.6 Examples.

Two examples are presented in this section.

The first example deals with some artificial data, which have been deliberately generated to challenge the method of analysis developed. Admittedly this example is far-fetched, but it does serve to illustrate several points of interest.

The real data of the second example concerned with detecting a possible shift in level of a chemical process variable, are taken from Box and Jenkins [1970].

### 2.6.1 An artificial example.

These data were generated by the model

$$y_i = \theta_1 x_{i,1} + \theta_2 x_{i,2} + e_i \qquad (2.102)$$

with $\theta_1 = 10.$ and $\theta_2 = 1.$, and where

$$\begin{cases} e_1 = M + \epsilon_1 \\ e_i = \phi e_{i-1} + \epsilon_i \end{cases} \qquad (2.103)$$

with $\phi = .5$, $M = 0$, and the $\epsilon_i$'s are i.i.d $N(0,1)$.

The values of the independent variable are listed in Table 2.I along with the synthetic "observations".



TABLE 2.I, The artificial data.

| i | $y_i$ | $x_{i,1}$ | $x_{i,2}$ |
|---|---|---|---|
| 1 | 181.436 | 1 | 81.920 |
| 2 | 140.528 | 1 | 40.960 |
| 3 | 119.204 | 1 | 20.480 |
| 4 | 110.097 | 1 | 10.240 |
| 5 | 104.308 | 1 | 5.120 |
| 6 | 101.009 | 1 | 2.560 |
| 7 | 100.490 | 1 | 1.280 |
| 8 | 99.487 | 1 | .640 |
| 9 | 98.130 | 1 | .320 |
| 10 | 98.970 | 1 | .160 |
| 11 | 99.696 | 1 | .080 |
| 12 | 102.054 | 1 | .040 |
| 13 | 100.001 | 1 | .020 |
| 14 | 100.644 | 1 | .010 |
| 15 | 98.955 | 1 | .005 |

The main purpose of this example is to illustrate a situation where the AR-1 process seems clearly stationary, but where the factors $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ and $|\underset{\approx}{\Xi}^{o\prime}\underset{\approx}{\Xi}^{o}|^{1/2}$ are very different in the priors (2.78) and (2.68) for the general and the stationary model respectively. It is then possible to see how this influences the marginal posterior distribution of $\phi$. Also this example demonstrates how the parametric inference about $\underset{\sim}{\theta}$ is affected by the presence of singularity points, of which this data set has two in relation to the general model but only one in relation to the stationary model.

Analyzing this data set on the basis of the general model it is seen, that not only is $\phi^* = 1$ a singularity point because of the mean $\theta_1$, but also $\phi^* = .5$ is a singularity point because of $\theta_2$ since

$\underset{\sim}{\Xi}_2 = \underset{\approx}{C}\,\underset{\sim}{X}_2$ vanishes in this point. This singularity may be explained by recognizing, that the exponential decay of $x_{i,2}$, i = 1,2,...,15, can be reproduced in the $e_i$'s by setting M = 81.92 and $\phi$ =.5.

How precisely $\underset{\sim}{\theta}$ is known given $\phi$, depends not only on $vs^2$ but also on the $\underset{\approx}{\Xi}'\underset{\approx}{\Xi}$-matrix in that point. Specifically the 100(1-$\alpha$)% highest posterior density (HPD) region for $\underset{\sim}{\theta}$ conditional on $\phi$ is proportional to $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{-1/2}(vs^2)^{1/2}$. The solid $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{-1/2}$-curve in Figure 2.5a shows how the overall precision with which $\underset{\sim}{\theta}$ may be known from the data degrades as $\phi$ approaches the singularity point $\phi^*$. In this example each singularity is attributed to a single linear parameter, and the two other curves in 2.5a show how the parameters individually become indeterminate as $\phi$ approaches their respective singularity points. $\sqrt{D_{11}}$ corresponds to $\theta_1$ and $\sqrt{D_{22}}$ to $\theta_2$, where $D_{jj}$ is defined by Equation (2.94). (All the curves in Figure 2.5a have been scaled to go through 1 at $\phi$ = 0).

Figure 2.5b shows posterior distributions for $\phi$. $p_u(\phi|\underset{\sim}{y})$ is the exact curve of Equation (2.96). It is noted, that this curve passes smoothly through $\phi$ = .5 and $\phi$ = 1., and that it can be well approximated by a t-distribution with n-p-2 degrees of freedom. The solid curve $p_n(\phi|\underset{\sim}{y})$ (2.101) has also the $t_{n-p-2}$ form, and is the marginal posterior using locally a noninformative prior as discussed in Section 2.5. Also the result of a direct multiplication of $p_u(\phi|\underset{\sim}{y})$ by $p_n(\phi)$ is shown, and it is seen, that for $\phi > 1$ where $p_u(\phi|\underset{\sim}{y})$ is essentially zero, the prior $p_n(\phi)$ begins to overpower the steadily decreasing likelihood and forces the posterior to diverge. This picture is typical for all examples considered, but is only drawn here in Figure 2.5b.



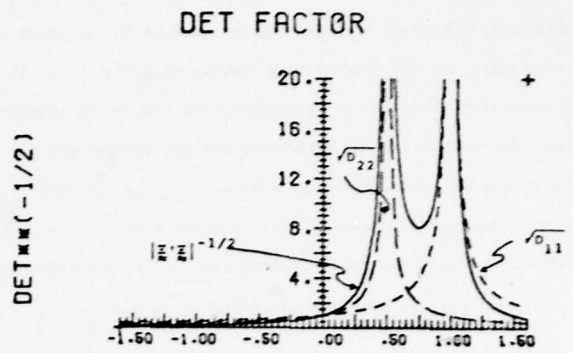



FIGURE 2.5a

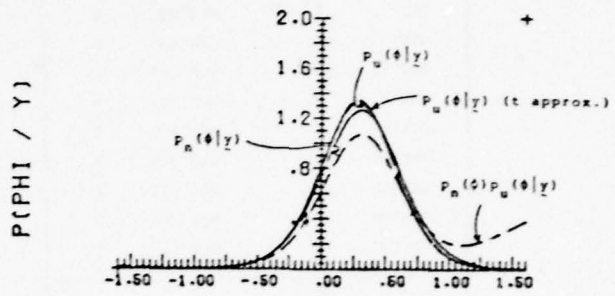



FIGURE 2.5b

Turning to the marginal inference about the linear parameters, the posterior density curves for $\theta_1$ and $\theta_2$ are drawn in Figure 2.6a and Figure 2.6b respectively. The solid curve in each figure is

$$p(\theta_j|\underset{\sim}{y}) = \int_R p(\theta_j|\phi,\underset{\sim}{y})\, p_n(\phi|\underset{\sim}{y})d\phi \qquad (2.104)$$

The curve $p_u(\theta_j|\underset{\sim}{y})$ is the posterior resulting from applying $p_u(\phi|\underset{\sim}{y})$ as a weight function in (2.104) instead of $p_n(\phi|\underset{\sim}{y})$. $p_n(\theta_j|\underset{\sim}{y})$ and $p(\theta_j|\underset{\sim}{y})$ are nearly coincidal for $\theta_1$ and indistinguishable for $\theta_2$.

For illustration $p_u(\theta_j|\underset{\sim}{y})$ has been approximated, as suggested in Section 2.4.3, by a weighted sum of 5 conditional t-distributions, $p(\theta_j|\phi,\underset{\sim}{y})$, specifically those corresponding to $\phi = -.8, -.4, .0, .4, .8$. This "5×t" approximation is very good indeed. It cannot be distinguished from $p_u(\theta_j|\underset{\sim}{y})$ for $j = 1$, and deviates only very slightly in the tail for $j = 2$.

Incidentally Figures 2.6a and 2.6b also show the consequences of an independence assumption, $p(\theta_j|\phi=0,\underset{\sim}{y})$; as well as the conditional inference on $\phi = \hat{\phi} = .29$. For both $\theta_1$ and $\theta_2$ the failure to recognize the serial correlation has the effect of overestimating the precision with which these parameters are known on the basis of the data, while in this example the location of the various curves are not much different. The approximation $p(\theta_j|\underset{\sim}{y}) \doteq p(\theta_j|\phi=\hat{\phi},\underset{\sim}{y})$ is good for $j = 1$, but not quite so good for $j = 2$. The probability mass lying outside the depicted interval $.85 < \theta_2 < 1.15$ is 15% for $p(\theta_2|\underset{\sim}{y})$ but only 2% for $p(\theta_2|\phi=\hat{\phi},\underset{\sim}{y})$.

On the basis of $p_n(\phi|\underset{\sim}{y})$ it may be concluded, that there is pretty firm evidence, that the model is stationary, $P(\phi < 1) = 98\%$; so if the further assumption of reversibility is made, then we may reanalyze the data on the basis of the stationary model.



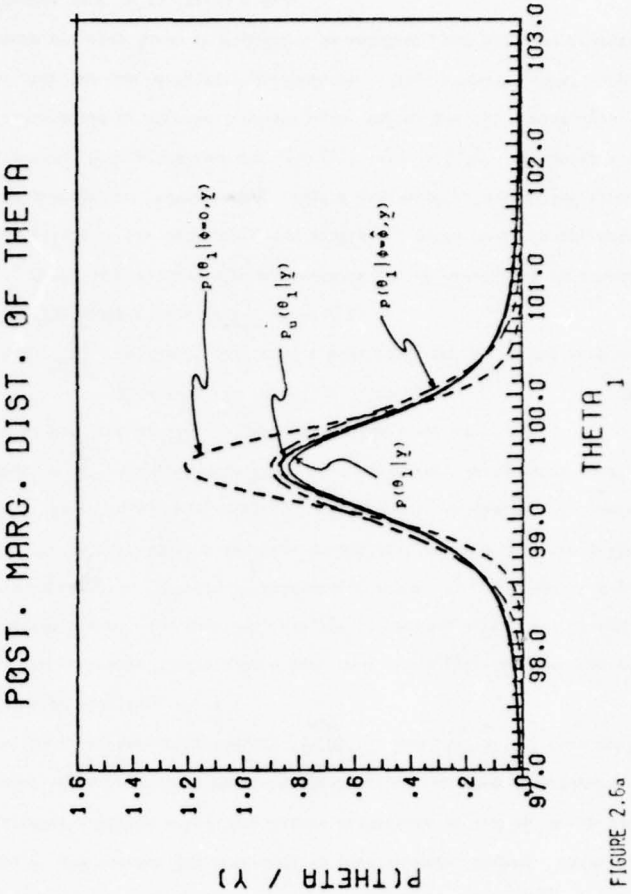



FIGURE 2.6a

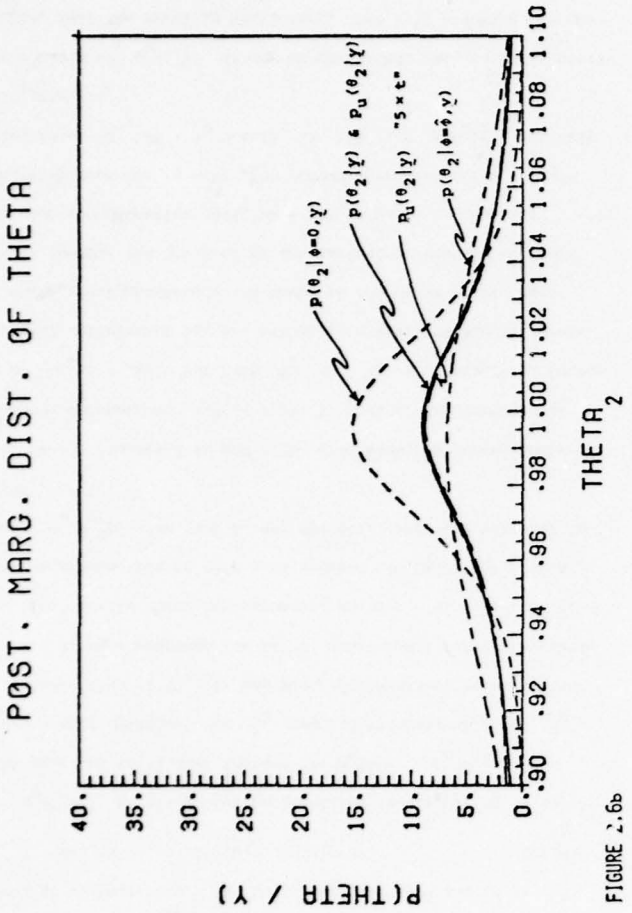



FIGURE 2.6b



*The marginal posterior distribution of* $\phi$ , $p^0(\phi|\underset{\sim}{y})$ (2.86), calculated on these premises is drawn in Figure 2.7b. Also $p_u(\phi|\underset{\sim}{y})$ and $p_n(\phi|\underset{\sim}{y})$ are redrawn for reference in this rescaled figure. Evidently this model revision makes very little difference as far as $\phi$ is concerned, which is particularly noteworthy since, as seen from Figures 2.5a and 2.7a, the prior factors $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ and $|\underset{\approx}{\Xi}^{0\prime}\underset{\approx}{\Xi}^{0}|^{1/2}$ are quite dissimilar functions of $\phi$ .

In contrast, Figure 2.8b shows, that $p^0(\theta_2|\underset{\sim}{y})$ is much more concentrated around $\theta_2$, than was $p(\theta_2|\underset{\sim}{y})$ in Figure 2.6b. The reason for this dramatic increase in information about $\theta_2$ is found in Figure 2.7a. It is seen that in relation to the stationary model, the point $\phi = .5$ is no longer a singularity point, since, unlike $\underset{\approx}{\Xi}$ , the second column of $\underset{\approx}{\Xi}^0$ does not degenerate to a vector of (n-1) zeros when $\phi = .5$, but (see (2.33)) to the n dimensional vector

$$\underset{\sim}{\Xi}_2^0 = (\sqrt{1-.5^2}\ \ 81.92,\ 0,\ldots,0)'. \qquad (2.105)$$

So now $\sqrt{D_{22}^0}$ is almost flat over a wide range of $\phi$ values (and substantially smaller than $\sqrt{D_{22}}$ near $\hat{\phi}$ ).

Under such unusual circumstances it may be beneficial to conduct a reanalysis of the data using the stationary model after having determined through the general model, that a stationarity assumption appears warranted. Ordinarily however $p^0(\theta_j|\underset{\sim}{y})$ and $p(\theta_j|\underset{\sim}{y})$ are quite alike as illustrated by Figures 2.6a and 2.8a where $p(\theta_1|\underset{\sim}{y}) \doteq p^0(\theta_1|\underset{\sim}{y})$, and similarly for the conditional posteriors. (Incidentally "5×t" approximations are also drawn in Figures 2.8a and 2.8b, and they are indistinguishable from $p^0(\theta_j|\underset{\sim}{y})$, j = 1,2.).

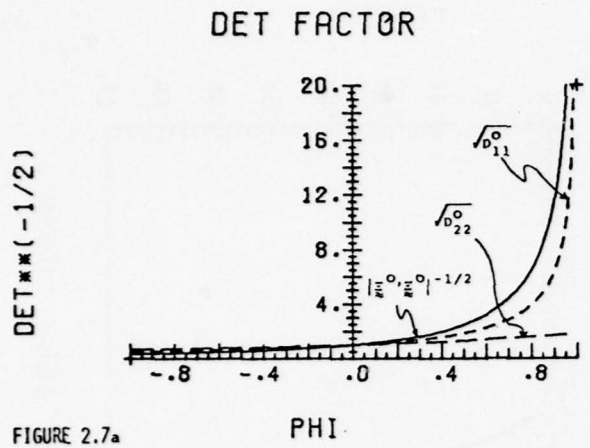



FIGURE 2.7a

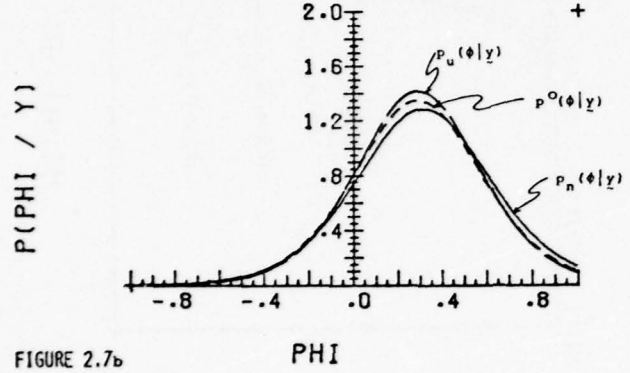



FIGURE 2.7b



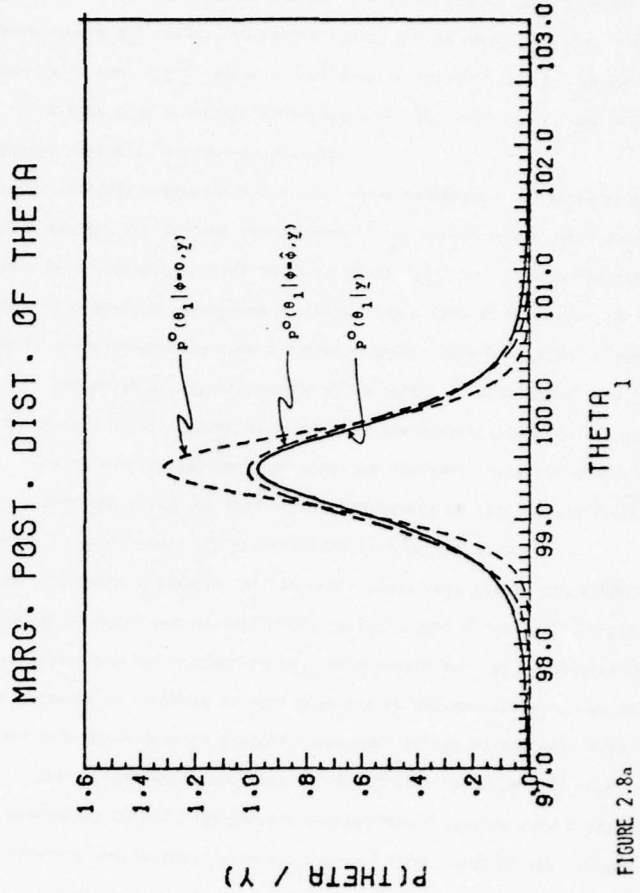



FIGURE 2.8a

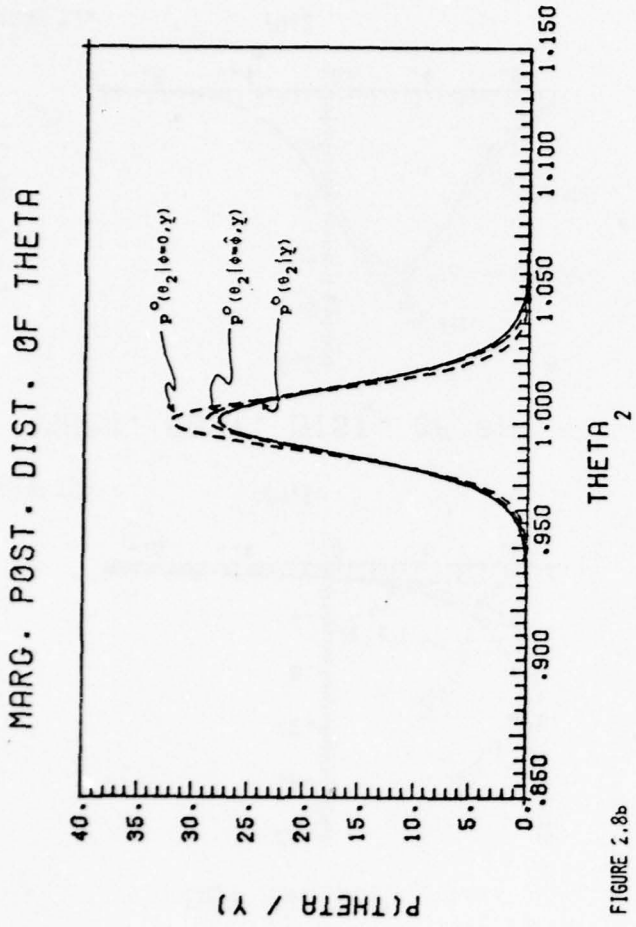



FIGURE 2.8b



2.6.2 Detecting a change in level.

The purpose of this example is to illustrate a situation where $\phi$ is close to one so that it is not entirely clear whether the process is stationary or not, and how the general model handles such a situation.

The 28 viscosity observations, $\underset{\sim}{y}$, listed in Table 2.II are taken from Box & Jenkins [1970], where they appear as the last column of their Series D. Suppose it were relevant to ask whether a change in level has taken place from the 22nd observation on. Then the data may be analyzed using the general model (2.16), with $\underset{\approx}{X}$ as given in Table 2.II; and where evidently $\theta_1$ is the process mean before the suspected change in level, and $\theta_2$ measures the size of the change.

The results of the analysis are displayed in Figures 2.9 and 2.10.

It is seen in Figure 2.9b, that the marginal posterior distribution of $\phi$, $p_u(\phi|\underset{\sim}{y})$ goes smoothly from stationary values to explosive ones. (Actually a t-approximation of $p_u(\phi|\underset{\sim}{y})$ is also drawn, but it cannot be distinguished from the exact curve). The posterior $p_n(\phi|\underset{\sim}{y})$ based on a locally noninformative prior has a tail area to the right of one with a probability mass equal to about 15%. So it is not entirely clear whether the process is stationary. Of course the general model allows marginal inference about the linear parameters $\underset{\sim}{\theta}$ to be made without having to assume stationarity.

Figure 2.9a shows the singularity of $\underset{\approx}{\Xi}'\underset{\approx}{\Xi}$ at $\phi = 1$ due to the mean $\theta_1$. The $\sqrt{D_{22}}$ curve shows, that in contrast to $\theta_1$ (with its associated $\sqrt{D_{11}}$ curve), inference about $\theta_2$ is apparently not very sensitive to $\phi$ in a region around $\phi = 1$, as far as spread is concerned. This indication is confirmed by Figures 2.10a and 2.10b showing posterior distributions of $\theta_1$ and $\theta_2$ respectively.

TABLE 2.II, The viscosity data.

| i | $y_i$ | $x_{i,1}$ | $x_{1,2}$ |
|---|---|---|---|
| 1 | 9.4 | 1 | 0 |
| 2 | 10.0 | 1 | 0 |
| 3 | 10.0 | 1 | 0 |
| 4 | 10.0 | 1 | 0 |
| 5 | 10.2 | 1 | 0 |
| 6 | 10.0 | 1 | 0 |
| 7 | 10.0 | 1 | 0 |
| 8 | 9.6 | 1 | 0 |
| 9 | 9.0 | 1 | 0 |
| 10 | 9.0 | 1 | 0 |
| 11 | 8.6 | 1 | 0 |
| 12 | 9.0 | 1 | 0 |
| 13 | 9.6 | 1 | 0 |
| 14 | 9.6 | 1 | 0 |
| 15 | 9.0 | 1 | 0 |
| 16 | 9.0 | 1 | 0 |
| 17 | 8.9 | 1 | 0 |
| 18 | 8.8 | 1 | 0 |
| 19 | 8.7 | 1 | 0 |
| 20 | 8.6 | 1 | 0 |
| 21 | 8.3 | 1 | 0 |
| 22 | 7.9 | 1 | 1 |
| 23 | 8.5 | 1 | 1 |
| 24 | 8.7 | 1 | 1 |
| 25 | 8.9 | 1 | 1 |
| 26 | 9.1 | 1 | 1 |
| 27 | 9.1 | 1 | 1 |
| 28 | 9.1 | 1 | 1 |



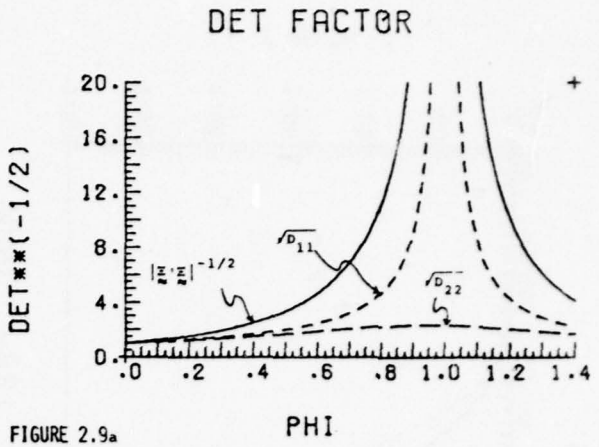



FIGURE 2.9a

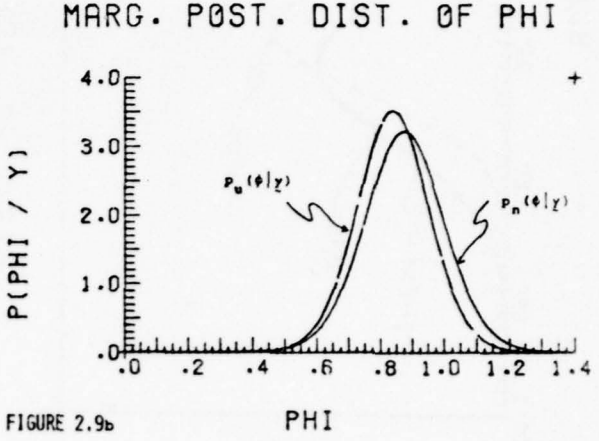



FIGURE 2.9b

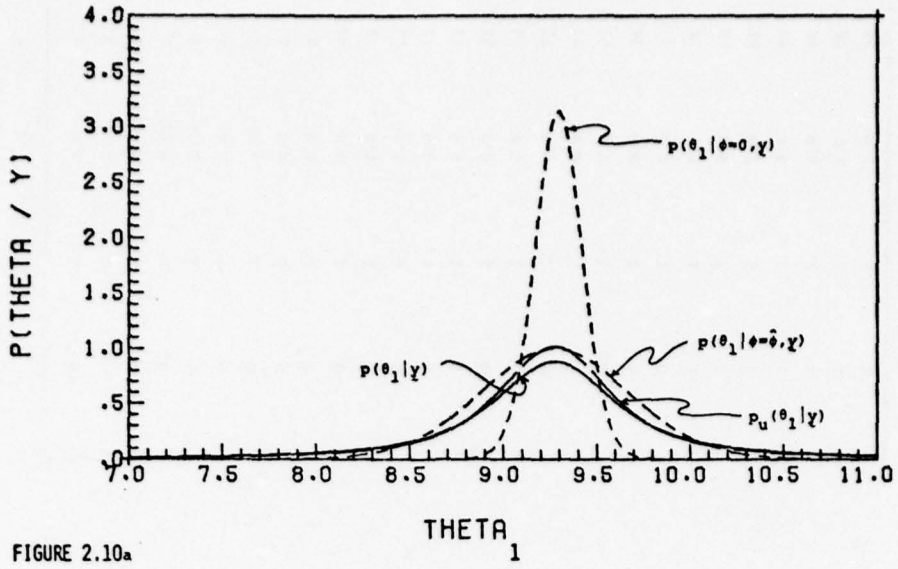



FIGURE 2.10a



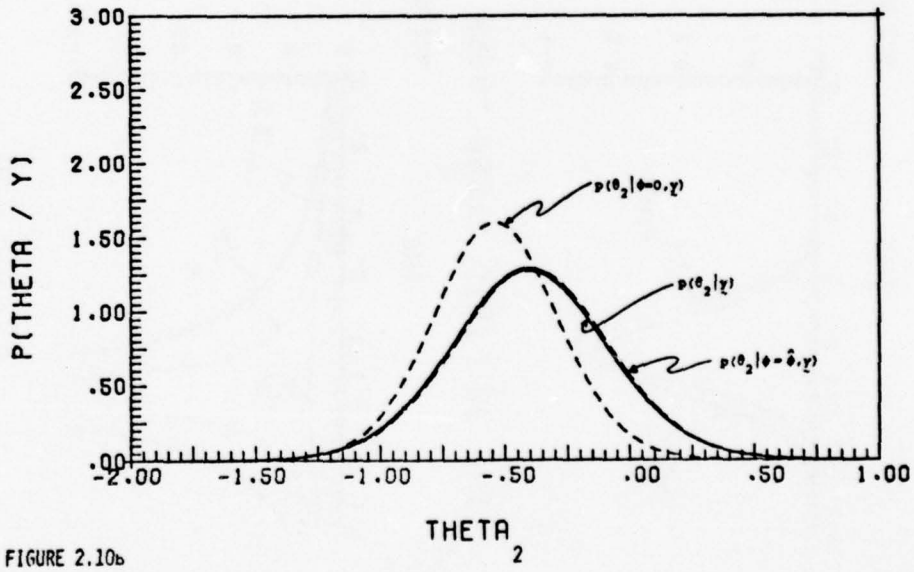



FIGURE 2.10b

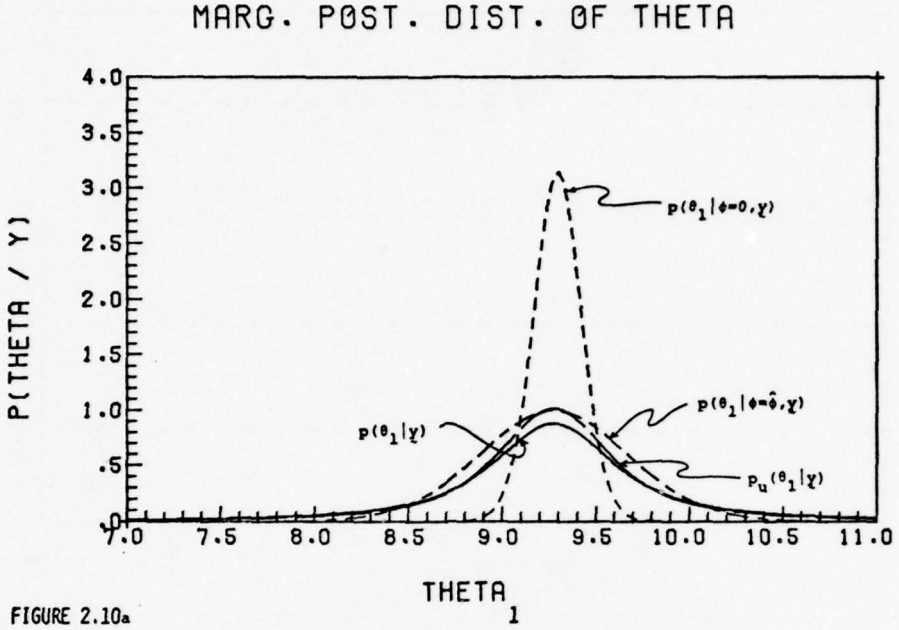



FIGURE 2.10a



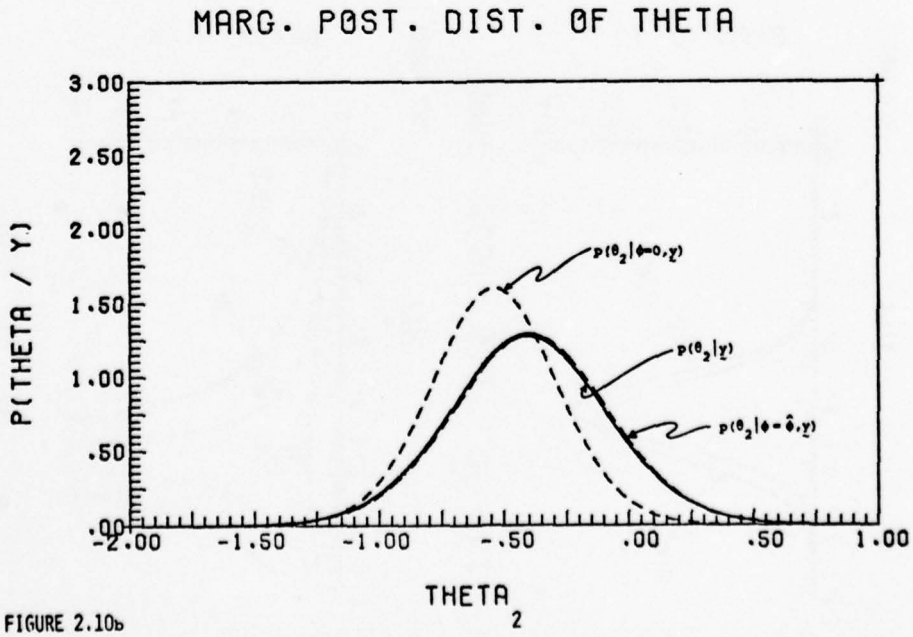



FIGURE 2.10b

The marginal posterior $p(\theta_2|\underset{\sim}{y})$ (2.104) in Figure 2.10b has a tail to the right of $\theta_2 = 0$ of 10%, hence the suggested level change is not particularly striking. Incidentally it is noted, that $p_u(\theta_2|\underset{\sim}{y})$, its "5×t" approximation, and $p(\theta_2|\phi=\hat{\phi}=.84,\underset{\sim}{y})$ are virtually identical to $p_n(\theta_2|\underset{\sim}{y})$, while $p(\theta_2|\phi=0,\underset{\sim}{y})$ gives a quite different picture ($P(0. < \theta_2) \doteq 2\%$).

Figure 2.10a shows similarly for the mean $\theta_1$, that $p_u(\theta_2|\underset{\sim}{y})$ and $p(\theta_2|\underset{\sim}{y})$ are much alike, and a "5×t" approximation of either curve produces a seemingly perfect match. Also $p(\theta_1|\phi=\hat{\phi}=.84,\underset{\sim}{y})$ represents a reasonably good approximation, while an independence assumption, $\phi = 0$, yields a very different resulting posterior $p(\theta_1|\phi=0,\underset{\sim}{y})$.

## 2.7 Conclusion.

The analysis of linear models with independent homoscedastic normal noise occupies a prominent position in statistics. If the assumption of independence cannot be made, it may often hold, that the dependence of the observations can adequately be accounted for by assuming that the noise in the data (or perhaps some transformation of the data, such as e.g. the logarithm or the first difference) follows a first order autoregressive (AR-1) scheme. In this chapter it was developed how such a widened class of linear models may be analyzed from a Bayesian point of view, specifically how inferences can be made about the linear parameters $\underset{\sim}{\theta}$ and the autoregressive parameter $\phi$ jointly, conditionally and marginally.

Two AR-1 schemes were considered. One covers (at a price of one additional starting parameter, M) explosive as well as stationary situations; the other assumes stationarity and reversibility a priori. Their corresponding likelihoods $\ell$ and $\ell^0$ were given, and complemented with



approximately noninformative priors to specify complete models. It was argued, that prior independence between $\underset{\sim}{\theta}$ and $\phi$ is not an appropriate assumption, in particular its adoption may lead to clearly unacceptable conclusions (Appendix B); and it was specifically suggested that $p(\underset{\sim}{\theta}|\phi)$ be set proportional to $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ ($|\underset{\approx}{\Xi}^{0\prime}\underset{\approx}{\Xi}^{0}|^{1/2}$ for the stationary model) where $\underset{\approx}{\Xi}$ ($\underset{\approx}{\Xi}^0$) is the matrix of transformed independent variables, which depends on $\phi$.

In situations where the data clearly have stationary noise the marginal posterior distribution for $\phi$, $p(\phi|\underset{\sim}{y})$, appears very little affected by which of the two AR-1 schemes is employed, even in a case where $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ and $|\underset{\approx}{\Xi}^{0\prime}\underset{\approx}{\Xi}^{0}|^{1/2}$ were drastically different. This incidentally supplies further support for the appropriateness of this factor in the prior. For truly stationary situations $\underset{\approx}{\Xi}'\underset{\approx}{\Xi}$ and $\underset{\approx}{\Xi}^{0\prime}\underset{\approx}{\Xi}^{0}$ will ordinarily be quite alike, so that little is to be gained from employing the a priori stationary model even in such cases. The not necessarily stationary model offers greater flexibility not only to let the data speak to the possibility of explosive noise, but also allows the disturbance associated with the first observation to be untypical. For this more general model, it was discovered, that special care has to be exercised in applying the locally approximately noninformative prior for $\phi$. A procedure for overcoming this difficulty was suggested, but as far as making inference about $\underset{\sim}{\theta}$ it appears from the examples, that it makes virtually no difference if instead a convenient locally uniform prior is utilized. In fact since prior information normally exists that $\phi$ cannot be much larger than 1 in practice, an informative prior like the uniform one may make good sense, because it concentrates the prior density more closely around $\phi = 0$ relative to a

noninformative prior. In any event the posterior density of $\phi$ derived from a uniform prior, can always be multiplied by any other prior to give its resulting posterior distribution.

The numerical integration leading to the marginal posterior distribution of $\underset{\sim}{\theta}$, $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ may be carried with a fine grid for $\phi$. It was demonstrated however, that excellent approximations result from computing $p(\theta_j|\underset{\sim}{y})$ as a weighted sum (the weights being proportional to $p(\phi|\underset{\sim}{y})$) of five t-distributions, $p(\theta_j|\phi,\underset{\sim}{y})$, corresponding to five equispaced $\phi$ values covering the interval where the body of $p(\phi|\underset{\sim}{y})$ is located.



Appendix A, On the noninformative prior for $\phi$.

In Section 2.3 approximately *noninformative* priors were produced on the basis of a data translating likelihood argument. As mentioned in Section 2.5 an alternative principle may consist of determining the prior $p(\psi)$ for $\psi$ which corresponds to a locally uniform prior for $\Psi$, where the $\Psi$-metric is such that the posterior density of $\Psi$ is approximately data translated.

In the following, this approach is outlined as it applies to the parameter $\phi$ in the general model. As it turns out, the very same (local) prior results, hence Figures A.1 through A.6 (plus 2.3 and 2.4) showing approximate data translated posteriors in the $\phi$-metric equally illustrate the effect of adopting the data translating likelihood principle.

From (2.85) we have (writing $SS(\phi)$ for $SS(\hat{\underset{\sim}{\theta}},\phi)$):

$$p(\phi|y) \propto (SS(\phi))^{-\frac{\nu+1}{2}} \qquad (A.1)$$

where $$\phi = \phi(\phi) = f(\phi) \qquad (A.2)$$

and $$\nu = n-p-2 \ . \qquad (A.3)$$

In this formulation, we shall determine $\phi = f^{-1}(\phi)$ such that the second derivative of (A.1), is approximately constant, in the point where the first derivative is zero. Actually $f^{-1}$ (or $f$) need not be determined; it is seen from

$$p(\phi) \propto \frac{d\phi}{d\phi} = \frac{df^{-1}(\phi)}{d\phi} \qquad (A.4)$$

that the first derivative of $f^{-1}$ (or of $f$) will suffice.

Because the posterior (A.1) is only specified up to a multiplicative constant, the log-posterior (lp for short) shall be employed, i.e. $p(\phi)$ is to be determined from

$$\left.\frac{\partial^2 \ell_p}{\partial \phi^2}\right|_{\frac{\partial \ell_p}{\partial \phi}=0} \doteq \text{constant} \qquad (A.5)$$

First

$$\frac{\partial \ell_p}{\partial \phi} = \frac{\partial \ell_p}{\partial \phi}\frac{d\phi}{d\phi} = -\frac{\nu+1}{2}\frac{\partial \ell n SS(\phi)}{\partial \phi} f'(\phi) \qquad (A.6)$$

Equating (A.6) to zero determines $\hat{\phi}$ and $\hat{\phi} = f^{-1}(\hat{\phi})$.

Next

$$\left.\frac{\partial^2 \ell_p}{\partial \phi^2}\right|_{\hat{\phi}} = -\frac{\nu+1}{2}\left.\frac{\partial^2 \ell n SS(\phi)}{\partial \phi^2}\left(\frac{d\phi}{d\phi}\right)^2\right|_{\hat{\phi}}$$

$$= -\frac{\nu+1}{2}\left(\frac{d\phi}{d\phi}\right)^2\left.\left[\frac{1}{SS(\phi)}\frac{\partial^2 SS(\phi)}{\partial \phi^2} - \left(\frac{\partial \ell n SS(\phi)}{\partial \phi}\right)^2\right]\right|_{\hat{\phi}}$$

$$= -\frac{\nu+1}{2}\left(\frac{d\phi}{d\phi}\right)^2 \frac{1}{SS(\hat{\phi})}\left.\frac{\partial^2 SS(\phi)}{\partial \phi^2}\right|_{\hat{\phi}}. \qquad (A.7)$$

Thus, to satisfy (A.5) it is required that

$$\left(\frac{d\phi}{d\phi}\right)^2 \propto \frac{1}{SS(\hat{\phi})}\left.\frac{\partial^2 SS(\phi)}{\partial \phi^2}\right|_{\hat{\phi}}. \qquad (A.8)$$

Since the right hand side of (A.8) not only depends on the data through $\hat{\phi}$, we shall settle for the lesser requirement, that the expected value of (A.5) is nearly constant. From

$$E\left(\frac{1}{SS(\hat{\phi})}\right) \doteq \frac{1}{\nu\sigma^2} \qquad (A.9)$$

and

$$E\left(\left.\frac{\partial^2 SS(\phi)}{\partial \phi^2}\right|_{\hat{\phi}}\right) = E\, 2\sum_{i=1}^{n-1} \hat{e}_i^2 \doteq 2E\sum_{i=1}^{n-1} e_i^2$$

$$= 2\sum_{i=1}^{n-1}\left(\phi^{2(i-1)}M^2 + \sum_{k=1}^{i}\phi^{2(k-1)}\sigma^2\right) \qquad (A.10)$$

we find



$$\left(\frac{\partial f^{-1}(\phi)}{\partial \phi}\right)^2 \propto \left(\frac{d\phi}{d\phi}\right)^2$$

$$\propto \frac{1}{(n-p-2)\sigma^2}\left(n\frac{\sigma^2}{1-\phi^2} - \frac{\sigma^2(1-\phi^{2n})}{(1-\phi^2)^2} + M^2\frac{1-\phi^{2(n-1)}}{1-\phi^2}\right)$$

$$\propto \frac{n}{n-p-2}\,\mathcal{J}(\phi) \qquad (A.11)$$

So as before we are led to the locally noninformative prior

$$p(\phi) \propto \left(\frac{1}{1-\phi^2} - \frac{1}{n}\frac{1-\phi^{2n}}{(1-\phi^2)^2}\right)^{1/2} \qquad (A.12)$$

Although $f^{-1}$ (or $f$) is not found analytically the data translating metric induced by the prior (A.12) can be constructed numerically. This is done in Figures A.1 through A.6 as well as in Figures 2.3 and 2.4. The "data" used in these 8 figures were generated as follows:

For Figures A.1, A.2, 2.3 and 2.4 the generating model is

$$y_i = e_i \qquad i = 1,2,\ldots,n \qquad (A.13)$$

with

$$\begin{cases} e_1 = \varepsilon_1 \\ e_i = \phi e_{i-1} + \varepsilon_i \end{cases} \qquad (A.14)$$

where the $\varepsilon_i$'s are i.i.d N(0,1). For the remaining figures the generating models are

Figure A.3: $y_i = \theta_1 + e_i$ with $\theta_1 = 1$. (A.15)

Figure A.4: $y_i = \theta_1 + \theta_2 \cdot i + e_i$ with $\theta_1 = 1.$ and $\theta_2 = 1.$ (A.16)

Figure A.5: $y_i = \theta_1 + \theta_2 \cdot 2^{14-i} + e_i$ with $\theta_1 = 1.$ and $\theta_2 = .01$ (A.17)

(i.e. a model like the one condisered in Section (2.6.1).

Figure A.6: $y_i = \theta_1 + \theta_2 x_i + e_i$ (A.18)

with $\theta_1 = 1.$ and $\theta_2 = 1.$

and where $x_i$ is a random walk $x_i = \sum_{j=1}^{i} a_j$,

the $a_j$'s being i.i.d. N(0,1)

(i.e. a model like Model A Equation (3.67) in Section 3.6 of Chapter 3.)

For all figures the employed sample size $n$ was 15, except for Figures 2.4 and A.2 where $n = 25$. (So differences between Figures 2.3 and B.1 are due to sampling variation only, and the same is true for Figure 2.4 vs. Figure A.2).

A reason for considering the model (A.13) is that the factor $|\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ in the prior vanished in this case, which perhaps makes the effect of the noninformative prior for $\phi$ more transparent. Also the posterior distribution of $\phi$, $p_e(\phi|\underset{\sim}{y})$, in relation to (A.13) is equal to the posterior of $\phi$ relating to any model

$$\underset{\sim}{y} = \underset{\approx}{X}\,\underset{\sim}{\theta} + \underset{\sim}{e} \tag{A.19}$$

conditional on $\underset{\sim}{\theta}$ :

$$p_e(\phi|\underset{\sim}{y}) \propto p_n(\phi)(SS(\phi))^{-\frac{n-1}{2}} \propto p_n(\phi)(\underset{\sim}{e}'\underset{\approx}{C}\,\underset{\sim}{e})^{-\frac{n-1}{2}} \propto p(\phi|\underset{\sim}{\theta},\underset{\sim}{y}) \tag{A.20}$$

For all the other models we have

$$p(\phi|\underset{\sim}{y}) \propto p_n(\phi)(SS(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p-1}{2}} \tag{A.21}$$

The various graphs in Figures 2.3, 2.4 (in Chapter 2) and A.1 through A.6 shall now be identified. Each figure depicts posterior distributions corresponding to five "data sets", all five generated from the same $\varepsilon_i$'s (different from figure to figure), but with



$\phi = -1.2, -.6, .0, .6, 1.2$. Looking at the "a"-figures relating to the original metric, the solid (nearly-) parabolas in the upper part of the figure are (scaled) SS-curves. Next the ∪-shaped prior $p_n(\phi)$ and the corresponding $q(\phi)$ factor are recognized (see Figure 2.2). The solid distribution curves $p_u(\phi|\underset{\sim}{y})$ are those resulting from adopting a locally uniform prior, while the broken curves are posteriors resulting from a locally noninformative prior. The $\phi$-metric induced by this prior is depicted in the "b"-figures. The divisions along the upper horizontal axis give a visual impression of how the original scale is streched and compressed relative to the $\phi$-metric. The five posterior distribution curves of $\phi$ are those given by (A.1).

ORIGINAL METRIC

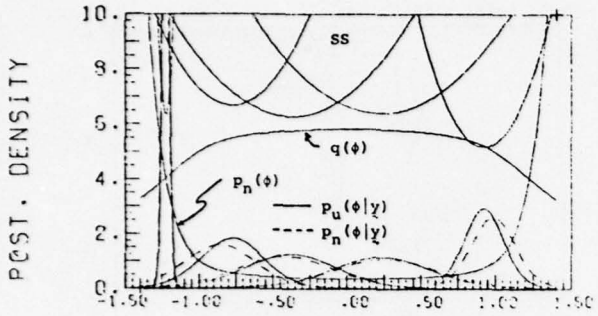



FIGURE A.1a

DATA TRANSLATING METRIC

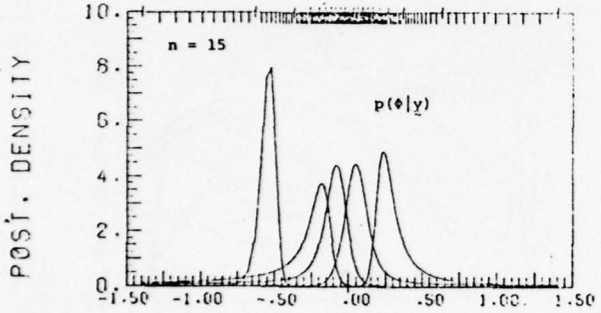



FIGURE A.1b



ORIGINAL METRIC

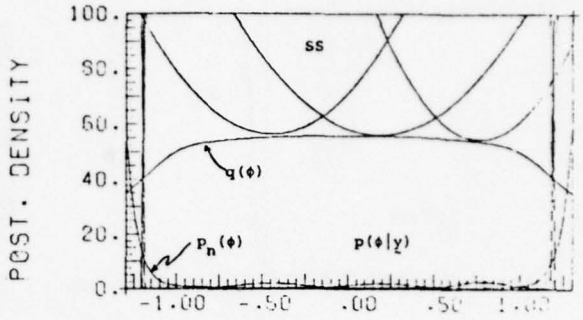



FIGURE A.2a

DATA TRANSLATING METRIC

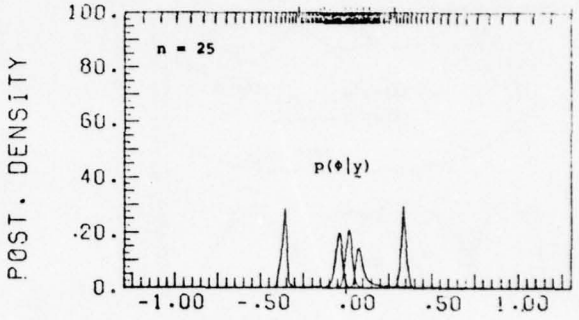



FIGURE A.2b

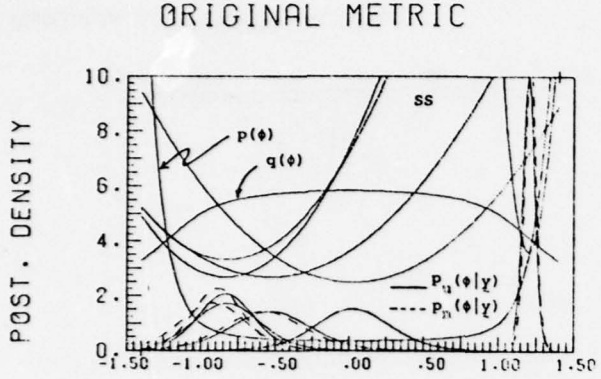



FIGURE A.3a

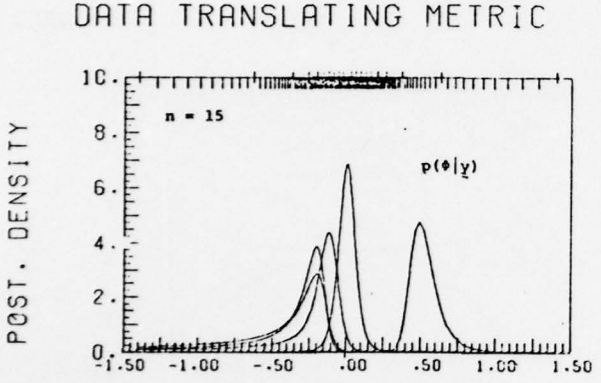



FIGURE A.3b



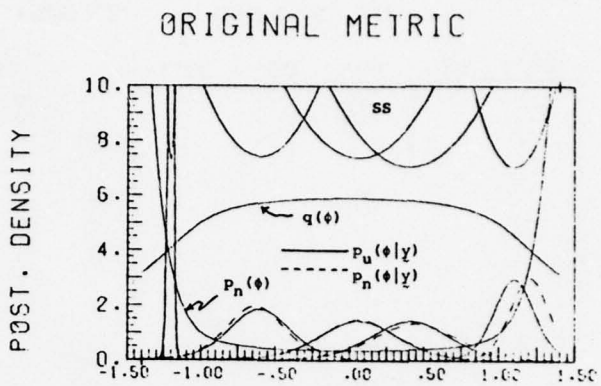



FIGURE A.4a

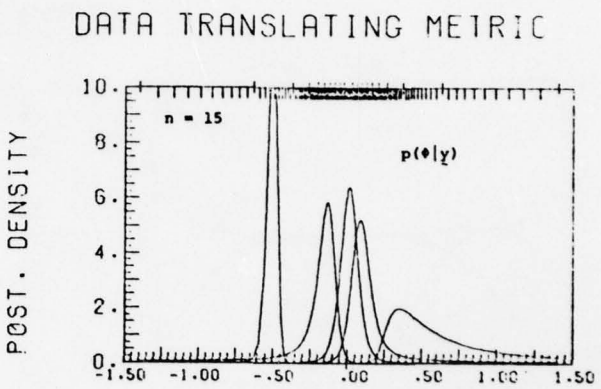



FIGURE A.4a

ORIGINAL METRIC

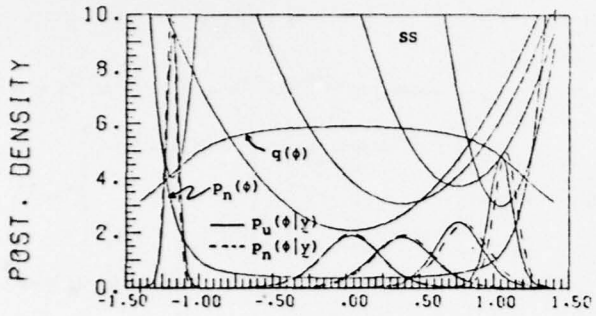



FIGURE A.5a

DATA TRANSLATING METRIC

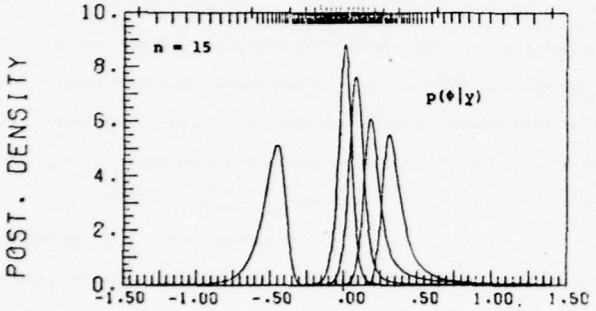



FIGURE A.5b



ORIGINAL METRIC

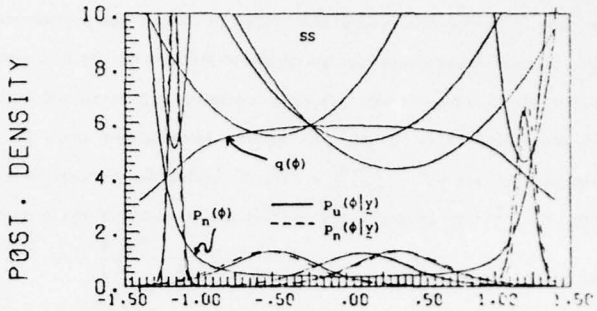



FIGURE A.6a

DATA TRANSLATING METRIC

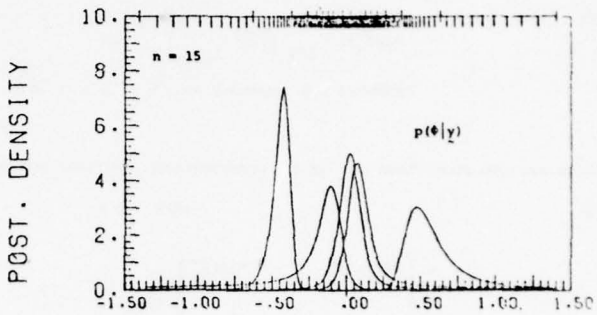



FIGURE A.6b

## Appendix B. Consequences of assuming prior independence between $\underset{\sim}{\theta}$ and $\phi$.

If independence is assumed between $\underset{\sim}{\theta}$ and $\phi$ a priori, then the factor $p(\underset{\sim}{\theta}|\phi)$ becomes a constant in the joint prior for the parameters $(M,\underset{\sim}{\theta},\phi,\sigma)$ of the general model. Consequently the marginal posterior distribution of $\phi$ takes the form

$$p_i(\phi|\underset{\sim}{y}) \propto |\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{-1/2}\, p(\phi|\underset{\sim}{y}) \tag{B.1}$$

where $p(\phi|\underset{\sim}{y})$ is the marginal posterior derived in Section 2.4 and given in Equation (2.85). The question we shall address here is how $p_i(\phi|\underset{\sim}{y})$ behaves in a neighbourhood of (the excluded) singularity points $\phi = \phi^*$ if any, where $\underset{\approx}{\Xi}$ has not full column rank, p. Writing

$$\underset{\approx}{\Xi} = [\underset{\sim}{\Xi}_1 \underset{\approx}{\Xi}(1)] \tag{B.2}$$

it may be assumed without loss of generality (see Section 3.2 of Chapter 3) that $\underset{\sim}{\Xi}_1 \to \underset{\sim}{0}$ for $\phi \to \phi^*$, while $[\underset{\sim}{X}_1{}^\dagger \underset{\approx}{\Xi}(1)]$ has full rank p in a neighbourhood of $\phi^*$, where $\underset{\sim}{X}_1{}^\dagger$ is the first column of $\underset{\approx}{X}$ excluding the last row.

From (2.27) we have

$$\underset{\sim}{\Xi}_1 = \begin{bmatrix} x_{2,1} - \phi\, x_{1,1} \\ \vdots \\ x_{n,1} - \phi\, x_{n-1,1} \end{bmatrix} = \begin{bmatrix} (\phi^* - \phi)\, x_{1,1} \\ \vdots \\ (\phi^* - \phi)\, x_{n-1,1} \end{bmatrix} = \delta \underset{\sim}{X}_1{}^\dagger \tag{B.3}$$

as $\phi^* = x_{2,1}/x_{1,1} = \ldots = x_{n,1}/x_{n-1,1}$, and where

$$\delta = \phi^* - \phi \tag{B.4}$$



Now

$$\begin{aligned} \det(\underset{\approx}{\Xi}'\underset{\approx}{\Xi}) &= \det([\delta \underset{\sim}{X}_1{}^\dagger \underset{\approx}{\Xi}(1)]'[\delta \underset{\sim}{X}_1{}^\dagger \underset{\approx}{\Xi}(1)]) \\ &= \delta^2 \det \begin{bmatrix} \underset{\sim}{X}_1{}^{\dagger\prime} \underset{\sim}{X}_1{}^\dagger & \underset{\sim}{X}_1{}^{\dagger\prime} \underset{\approx}{\Xi}(1) \\ \underset{\approx}{\Xi}'(1) \underset{\sim}{X}_1{}^\dagger & \underset{\approx}{\Xi}'(1) \underset{\approx}{\Xi}(1) \end{bmatrix} \\ &= \delta^2 \; d(\phi) \end{aligned} \tag{B.5}$$

where the function (determinant) $d(\phi)$ is continuous and not zero at $\phi = \phi^*$.

For $0 < \varepsilon_1 < \varepsilon_2$ we consider the integral

$$\begin{aligned} &\lim_{\varepsilon_1 \to 0} \int_{\phi^*-\varepsilon_2}^{\phi^*-\varepsilon_1} |\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{-1/2}\, p(\phi|\underset{\sim}{y}) d\phi \\ &= \lim_{\varepsilon_1 \to 0} \left( d(\phi_d)^{-1/2}\, p(\phi = \phi_d|\underset{\sim}{y}) \int_{\phi^*-\varepsilon_2}^{\phi^*-\varepsilon_1} \frac{1}{\delta}\, d\phi \right) \end{aligned} \tag{B.6}$$

where for sufficiently small $\varepsilon_2$ this equality holds for some $\phi_d \in (\phi^*-\varepsilon_2, \phi^*)$. Now since

$$\int_{\phi^*-\varepsilon_2}^{\phi^*-\varepsilon_1} \frac{1}{\delta}\, d\phi = \ln \frac{\varepsilon_2}{\varepsilon_1} \tag{B.7}$$

obviously the integral in (B.6) does not converge as $\varepsilon_1 \to 0$. This implies, that if the factor $p(\underset{\sim}{\theta}|\phi) = |\underset{\approx}{\Xi}'\underset{\approx}{\Xi}|^{1/2}$ is left out of the prior, i.e. if prior independence between $\underset{\sim}{\theta}$ and $\phi$ is assumed, then the resulting posterior distribution $p_i(\phi|\underset{\sim}{y})$ concentrates all posterior probability in the neighbourhood(s) of the singularity point(s) if any, regardless of the data. In particular models containing a mean have such a singularity point namely $\phi^* = 1$.

This clearly unacceptable consequence of adopting $p(\underset{\sim}{\theta}|\phi) = p(\underset{\sim}{\theta}) \propto k$ as a prior for $\underset{\sim}{\theta}$ in the general model, of course also arises in the context of the stationary model.

## CHAPTER 3

### Probing for Serial Correlation in Least Squares Regression.

Writing the regression model as

$$\underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times p}{\underset{\approx}{X}}\ \underset{p\times 1}{\underset{\sim}{\theta}} + \underset{n\times 1}{\underset{\sim}{e}}, \qquad (3.1)$$

where $\underset{\sim}{y}$ is the vector of observations, $\underset{\approx}{X}$ is the matrix of independent variables, and $\underset{\sim}{\theta}$ is the vector of the $p$ linear parameters; it is usually assumed, that the random errors $e_i$, $i = 1,2,\ldots,n$, are independently distributed as $N(0,\sigma^2)$, and this assumption provides justification for the "standard" techniques of analysis.

For data collected in time (or space) sequence it may often be true, that the errors are in fact serially dependent, in particular this would seem likely, for many economic and business series.

Testing for serial correlation in situations of this kind, has been given much attention by a number of authors, most notably by Durbin and Watson [1950], [1951], [1971], whose well known test (DW-test for short) enjoys widespread use in applied work in many fields. In testing the null hypothesis of independence, the DW-statistic is constructed with a first order autoregressive (AR-1) noise structure (3.2) as the alternative hypothesis in mind.

$$e_i = \phi e_{i-1} + \varepsilon_i \qquad (3.2)$$

where the $\varepsilon_i$'s are i.i.d. $N(0,\sigma^2)$, i.e. they are a white noise series (also independent of $\underset{\approx}{X}$).

Modifications and new tests have been proposed from time to time, for example by Abrahamse and Louter [1971], Berenblut and Webb [1973], Durbin [1969], [1970], Grary [1970] and Schmidt [1972]. From the point of view of power, however, none of these alternatives have been



demonstrated to have any decisive advantage over the DW-test, not even when the true noise structure is different from AR-1 and the competing test was devised for such situations, Smith [1976] (also Abrahamse and Lauter [1971], Berenblut and Webb [1973], Durbin and Watson [1971]).

The whole idea of testing for serial correlation in the circumstances in which it is usually done is somewhat suspect. For sequential data such as is often obtained in economics for example, one would usually expect serial correlation. There is therefore no particular reason for believing the null hypothesis, and it would often seem much more natural to assume, that the noise vector $\underset{\sim}{e}$ in (3.1) is generated by an AR-1 scheme like (3.2), rather than by a white noise process.

It is the topic of this chapter to see how a broader appreciation of the inferential situation concerning an autoregressive parameter $\phi$ may be achieved, from a likelihood as well as a Bayesian point of view. In particular when $\phi$ behaves as a nuisance parameter in relation to $\underset{\sim}{\theta}$ in (3.1), to consider how inferences about $\underset{\sim}{\theta}$ may be made taking $\phi$ into account.

In cases where the null hypothesis has meaning, the likelihood and the Bayesian approaches suggest alternative testing procedures, which can be viewed as competitors to the DW-test. A simulation study is undertaken to determine their relative powers in conjunction with two selected model structures, as well as to shed light on other aspects of the alternatives.

Three examples are presented to illustrate consequences of employing different approaches.

## 3.1 The Durbin-Watson test.

Durbin and Watson's test statistic is defined as

$$d = \frac{\sum_{i=1}^{n-1} (r_{i+1}-r_i)^2}{\sum_{i=1}^{n} r_i^2} \qquad (3.3)$$

where the $r_i$'s are the residuals after an ordinary least squares fit, i.e.

$$\underset{\sim}{r} = (I - \underset{\approx}{X}(\underset{\approx}{X}'\underset{\approx}{X})^{-1}\underset{\approx}{X}')\underset{\sim}{y} \qquad (3.4)$$

and it is assumed without loss of generality that $\underset{\approx}{X}$ has full column rank, p. The statistic d takes values between 0 and 4. Values close to 2 imply that the $e_i$'s in the linear model (3.1) are independent, while positive serial correlation will tend to produce smaller values and negative serial correlation larger ones. Unfortunately the sampling distribution of d under the null hypothesis of independence depends in general on $\underset{\approx}{X}$, so no single significance point $d_S(\alpha)$ depending only on n and p can be given. But bounds have been found $d_L(\alpha) \leq d_S(\alpha) \leq d_U(\alpha)$. The one-sided DW-test for serial correlation at the $\alpha$ level, thus consists of referring d to these tabulated bounds. If $d < d_L(\alpha)$ significance has been established, i.e. the hypothesis is to be rejected at the level $\alpha$. If $d_L(\alpha) < d < d_U(\alpha)$ the test is said to be inconclusive; and employing popular terminology, the null hypothesis is to be accepted when $d_U(\alpha) < d$, although no more than lack of significance has been established at the given level.

Testing for negative serial correlation is done by referring the statistic (4-d) to the same bounds, and a two tail test at the $2\alpha$ level results from doing both. The test would ordinarily be credited with the power associated with $d_L(\alpha)$ only; an exception is polynomial regression



where it has been shown that $d_U$ gives a very good approximation to the true significance points of d, Hannan [1957].

Rewriting (3.3) slightly as

$$d = \frac{\sum_{i=2}^{n} r_i^2 + \sum_{i=1}^{n-1} r_i^2 - 2\sum_{i=1}^{n-1} r_i r_{i+1}}{\sum_{i=1}^{n} r_i^2} \doteq 2 - 2\frac{\sum_{i=1}^{n-1} r_i r_{i+1}}{\sum_{i=1}^{n} r_i^2} \qquad (3.5)$$

it is recognized that the fraction on the far right hand side is of the same form as (2.49) and (2.44), and hence provides an estimate of $\phi$ which is essentially equivalent to performing only one iteration in the scheme of Figure 2.1 in Chapter 2. Denoting this estimate by $\check{\phi}$, i.e.

$$\check{\phi} = 1 - \frac{d}{2} \qquad (3.6)$$

the DW-test might be formulated in terms of $\check{\phi}$, in which case the test consists of determining whether $\check{\phi}$ is significantly different from zero. This formulation is perhaps somewhat preferable, because the corresponding bounds help to give a direct indication of how pronounced the serial correlation must be in order to be detected. For example if n = 15 and p = 2, we find $\phi_L(.05) = 1 - d_L(.05)/2 = .46$; implying that autocorrelation caused by a true value of say .3, might very well pass unnoticed.

The DW-test presupposes, that the linear regression model under consideration includes a mean. Durbin and Watson argue [1950], that this is only a slight limitation since a model lacking a mean can be augmented to include one for the purpose of conducting the test, and subsequently this extra independent variable may be eliminated again. Cochrane and Orcutt [1949] have shown, that the residuals $\underset{\sim}{r}$ from an ordinary least squares fit are biased towards randomness, and increasingly so as the number of independent variables is expanded. This

result would lead one to infer, contrary to Durbin and Watson, that the evidence of serial correlation carried by $\underset{\sim}{r}$, would tend to be obscured by forcing a mean into the linear model. Such an effect is indeed demonstrated in Sections 3.5 and 3.6.

Durbin and Watson do not address the question of what to do if the null-hypothesis is rejected. Cochrane and Orcutt [1949] demonstrated that the ill effects of ignoring positive serial correlation of the errors in least squares regression might be remedied by applying an independence inducing transformation, and analyzing the transformed linear model by "standard" methods. If the error follows an AR-1 scheme, independence is induced by multiplying (3.1) by the matrix $\underset{\approx}{c}$:

$$\underset{\approx}{c}\,\underset{\sim}{y} = \underset{\approx}{c}\,\underset{\approx}{X}\,\underset{\sim}{\theta} + \underset{\approx}{c}\,\underset{\sim}{e} \tag{3.7}$$

$$\underset{(n-1)\times 1}{\underset{\sim}{z}} = \underset{(n-1)\times p}{\underset{\approx}{\Xi}}\;\underset{p\times 1}{\underset{\sim}{\theta}} + \underset{(n-1)\times 1}{\underset{\sim}{\varepsilon}} \tag{3.8}$$

where

$$\underset{\approx}{c} = \begin{bmatrix} -\phi & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -\phi & 1 \end{bmatrix} \tag{3.9}$$

Theil and Nagar [1961] suggest to proceed as if $\underset{\approx}{c}$ was known and was given by substituting $\check{\phi}$ for $\phi$ in (3.9) (actually they employ a slight variant of $\check{\phi}$). They caution, that not only does the approximation $\phi = \check{\phi}$ neglect sampling variation in $\check{\phi}$; but also that the same data are used to test the null hypothesis of independence, to estimate $\phi$ if this hypothesis is rejected, and to reestimate $\underset{\sim}{\theta}$ in that case. Hence they conclude, that the usual standard deviations for the least squares estimates of the $\theta_j$'s in (3.8) may, on the average, underestimate the unreliability of these estimates.



## 3.2 $SS(\hat{\underset{\sim}{\theta}},\phi)$ as a function of $\phi$.

*Before turning to the likelihood and Bayesian treatment, it is* important to study $SS(\hat{\underset{\sim}{\theta}},\phi)$ as a function of $\phi$, where $SS(\hat{\underset{\sim}{\theta}},\phi)$ is the residual sum of squares from fitting by least squares the transformed regression model (3.8), i.e.

$$SS(\hat{\underset{\sim}{\theta}},\phi) = \underset{\sim}{z}'(\underset{\approx}{I} - \underset{\approx}{\Xi}(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}')\underset{\sim}{z} \tag{3.10}$$

It is assumed without loss of generality that $\underset{\approx}{X}$ has full column rank, p; however $\underset{\approx}{\Xi} = \underset{\approx}{c}\,\underset{\approx}{X}$ need not have full column rank for all $\phi$. It is therefore of interest to see how $SS(\hat{\underset{\sim}{\theta}},\phi)$ behaves when $\phi$ passes through singularity points (if any).

### 3.2.1 Singularity due to a mean.

If the original linear model (3.1) involves a mean, say $\theta_1$, so that

$$\underset{\approx}{X} = [\,\underset{\sim}{1}\;\underset{\approx}{X}_{(1)}\,] \tag{3.11}$$

where $\underset{\sim}{1}$ is a vector of ones, then

$$\underset{\approx}{\Xi} = \underset{\approx}{c}\,\underset{\approx}{X} = \begin{bmatrix} 1-\phi & \\ 1-\phi & \\ \vdots & \underset{\approx}{c}\,\underset{\approx}{X}_{(1)} \\ 1-\phi & \end{bmatrix} = \begin{bmatrix} \delta & \\ \underset{\sim}{\delta} & \underset{\approx}{\Xi}_{(1)} \end{bmatrix} \tag{3.12}$$

where $\underset{\sim}{\delta} = \delta\,\underset{\sim}{1} = (1-\phi)\underset{\sim}{1}$.

Clearly the transformed model (3.8) has less than full rank at the point $\phi = 1$ ($\delta = 0$), and the matrix $\underset{\approx}{\Xi}'\underset{\approx}{\Xi}$ becomes singular. The question is how $SS(\hat{\underset{\sim}{\theta}},\phi)$ behaves as $\phi \to 1$.

Restricting attention to a neighbourhood around $\phi = 1$, where $\underset{\approx}{\Xi}_{(1)}$ has full rank equal to $p - 1$ (which is a proper restriction as proven below), we first write

$$\underset{\approx}{\Xi}'\underset{\approx}{\Xi} = [\underset{\sim}{\delta}\ \underset{\approx}{\Xi}_{(1)}]'[\underset{\sim}{\delta}\ \underset{\approx}{\Xi}_{(1)}] = \begin{bmatrix} \underset{\sim}{\delta}'\underset{\sim}{\delta} & \underset{\sim}{\delta}'\ \underset{\approx}{\Xi}_{(1)} \\ \underset{\approx}{\Xi}'_{(1)}\underset{\sim}{\delta} & \underset{\approx}{\Xi}'_{(1)}\underset{\approx}{\Xi}_{(1)} \end{bmatrix}. \tag{3.13}$$

Then using the well known formula for the inverse of a partitioned matrix we have

$$(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1} = \begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \underset{\approx}{E} \end{bmatrix} + F \begin{bmatrix} 1 & -\underset{\sim}{\delta}'\underset{\sim}{\Xi}_{(1)}\underset{\approx}{E} \\ -\underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{\delta} & \underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{\delta}\underset{\sim}{\delta}'\underset{\approx}{\Xi}_{(1)}\underset{\approx}{E} \end{bmatrix} \tag{3.14}$$

with $\underset{\approx}{E} = (\underset{\approx}{\Xi}'_{(1)}\ \underset{\approx}{\Xi}_{(1)})^{-1}$, and where

$$F = [\underset{\sim}{\delta}'\underset{\sim}{\delta} - \underset{\sim}{\delta}'\underset{\approx}{\Xi}_{(1)}\ \underset{\approx}{E}\ \underset{\approx}{\Xi}'_{(1)}\ \underset{\sim}{\delta}]^{-1} \tag{3.15}$$

$$= \delta^{-2}[\underset{\sim}{1}'\underset{\sim}{1} - \underset{\sim}{1}'\underset{\approx}{\Xi}_{(1)}\ \underset{\approx}{E}\ \underset{\approx}{\Xi}'_{(1)}\ \underset{\sim}{1}]^{-1} = \delta^{-2}f$$

Now for any $\delta$ in a neighbourhood around $0$, excluding that point, the identity holds:

$$\begin{aligned} &\underset{\approx}{\Xi}(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\ \underset{\approx}{\Xi}' \\ &= [\underset{\sim}{\delta}\ \underset{\approx}{\Xi}_{(1)}]\left\{\begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \underset{\approx}{E} \end{bmatrix} + f\begin{bmatrix} \delta^{-2} & -\delta^{-1}\underset{\sim}{1}'\underset{\approx}{\Xi}_{(1)}\ \underset{\approx}{E} \\ -\underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{1}\delta^{-1} & \underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{1}\underset{\sim}{1}'\underset{\approx}{\Xi}_{(1)}\underset{\approx}{E} \end{bmatrix}\right\}[\delta\ \underset{\approx}{\Xi}_{(1)}]' \\ &= [\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]\left\{\begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \underset{\approx}{E} \end{bmatrix} + f\begin{bmatrix} 1 & -\underset{\sim}{1}'\ \underset{\approx}{\Xi}_{(1)}\underset{\approx}{E} \\ -\underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{1} & \underset{\approx}{E}\,\underset{\approx}{\Xi}'_{(1)}\underset{\sim}{1}\underset{\sim}{1}'\ \underset{\approx}{\Xi}_{(1)}\ \underset{\approx}{E} \end{bmatrix}\right\}[\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]' \\ &= [\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]\ ([\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]'[\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}])^{-1}[\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]'\ . \end{aligned} \tag{3.16}$$

The last step in (3.16) requires that the matrix $[\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]$ is nonsingular, i.e. that $\underset{\sim}{1}$ does not lie in the vector space spanned by the columns of $\underset{\approx}{\Xi}_{(1)}, \underset{\sim}{\Xi}_2\ \underset{\sim}{\Xi}_3\ \cdots\ \underset{\sim}{\Xi}_p$, when $\phi = 1$. The next to last expression in (3.16) does not involve $\delta$ and hence $SS(\hat{\underset{\sim}{\theta}},\phi)$ has a limiting value as $\phi \to 1$. If we now define



$$SS(\hat{\underset{\sim}{\theta}},\phi = 1) = \lim_{\phi \to 1} SS(\hat{\underset{\sim}{\theta}},\phi) \tag{3.17}$$

then $SS(\hat{\underset{\sim}{\theta}},\phi)$ goes continuously through the singularity point $\phi = 1$.

When the last transcription in (3.16) is valid, then the limiting value (3.17) may be calculated conveniently from (3.10) by replacing the first (vanishing) column of $\underset{\approx}{\Xi}$, evaluated at $\phi = 1$, by a column of ones. It is recognized, that this quantity is a residual sum of squares corresponding to a least squares fit of the model (3.8) for $\phi = 1$, where the original mean, $\theta_1$, has been replaced by a new parameter $\theta_1^*$ which has the interpretation of a mean for the differences $z_2 = y_2 - y_1$, $z_3 = y_3 - y_2, \ldots, z_n = y_n - y_{n-1}$.

3.2.2 Simple singularity.

In slightly greater generality there may exist other values of $\phi, \phi^*$ say, different from unity, causing some column $\underset{\sim}{\Xi}_j$ to vanish, i.e.

$$\underset{\sim}{\Xi}_j \to \underset{\sim}{0} \quad \text{as} \quad \phi \to \phi^*. \tag{3.18}$$

It is convenient to introduce the symbols $\underset{\sim}{X}_j^\dagger$ and $\underset{\sim}{X}_j^\dagger$ whose meaning is explained by

$$\underset{\sim}{X}_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_j^\dagger \\ x_{n,j} \end{bmatrix} = \begin{bmatrix} x_{1,j} \\ \underset{\sim}{X}_j^\dagger \end{bmatrix}. \tag{3.19}$$

Clearly

$$\underset{\sim}{\Xi}_j = \underset{\sim}{X}_j^\dagger - \phi\ \underset{\sim}{X}_j^\dagger \tag{3.20}$$

so that

$$\begin{aligned} &\underset{\sim}{\Xi}_j = \underset{\sim}{0} \Leftrightarrow \underset{\sim}{X}_j^\dagger = \phi^*\ \underset{\sim}{X}_j^\dagger \\ &\Leftrightarrow x_{i,j} = \phi^* x_{i-1,j} \quad i = 2,3,\ldots,n \\ &\Leftrightarrow \phi^* = x_{2,j}/x_{1,j} = \cdots = x_{n,j}/x_{n-1,j} \end{aligned} \tag{3.21}$$

and we may write

$$\underset{\sim}{\Xi}_j = \begin{bmatrix} x_{2,j} - \phi x_{1,j} \\ \vdots \\ x_{n,j} - \phi x_{n-1,j} \end{bmatrix} = \begin{bmatrix} (\phi^*-\phi)x_{1,j} \\ \vdots \\ (\phi^*-\phi)x_{n-1,j} \end{bmatrix} = (\phi^*-\phi)\underset{\sim}{X}_j^\dagger = \delta \underset{\sim}{X}_j^\dagger \quad (3.22)$$

where $\delta = \phi^* - \phi$ .

The problem is now of exactly the same form as above in subsection 3.2.1. Thus the conclusion may be drawn as before, that $SS(\hat{\underset{\sim}{\theta}},\phi)$ has a limit for $\phi \to \phi^*$, so that $SS(\hat{\underset{\sim}{\theta}},\phi)$ becomes continuous through $\phi = \phi^*$ by defining

$$SS(\hat{\underset{\sim}{\theta}},\phi^*) = \lim_{\phi \to \phi^*} SS(\hat{\underset{\sim}{\theta}},\phi) \ . \quad (3.23)$$

Again, provided $\underset{\sim}{X}_j^\dagger$ is linearly independent of $\underset{\sim}{\Xi}_1 \cdots \underset{\sim}{\Xi}_{j-1}, \underset{\sim}{\Xi}_{j+1} \cdots \underset{\sim}{\Xi}_p$ when $\phi = \phi^*$, the limiting value (3.23) may be interpreted as a residual sum of squares from fitting the linear model (3.8) with $\phi = \phi^*$, where a new parameter $\theta_j^*$ has been substituted for the one that became redundant for $\phi = \phi^*$ .

It is noted, that at most one column may vanish for a particular value $\phi^*$ of $\phi$ . It is seen from (3.21) that $\phi^*$ determines $\underset{\sim}{X}_j$ up to a multiplication constant, so that the occurance of more than one column vanishing for $\phi = \phi^*$ is contrary to the assumption of $\underset{\approx}{X}$ having full column rank.

3.2.3 *General singularity.*

In general, singularity may occur at $\phi = \phi^*$ when

$$k_1 \underset{\sim}{\Xi}_1 + k_2 \underset{\sim}{\Xi}_2 + \ldots + k_p \underset{\sim}{\Xi}_p = \underset{\sim}{0} \ . \quad (3.24)$$

Assuming without loss of generality that $k_1 = -1$ we have equivalently

$$\underset{\sim}{X}_1^\dagger - \phi^* \underset{\sim}{X}_1^\dagger = k_2(\underset{\sim}{X}_2^\dagger - \phi^* \underset{\sim}{X}_2^\dagger) + \ldots + k_p(\underset{\sim}{X}_p^\dagger - \phi^* \underset{\sim}{X}_p^\dagger). \quad (3.25)$$

This expression shows, that a situation of singularity cannot persist over a range of $\phi$ values. (3.25) is equating two first order polynomials in $\phi$ (n-1 times over), and can only be satisfied throughout an interval if the (n-1 sets of) coefficients are identical, i.e.



$$\begin{cases} \underset{\sim}{X}_1^\dagger = k_2 \underset{\sim}{X}_2^\dagger + \ldots + k_p \underset{\sim}{X}_p^\dagger \\ \underset{\sim}{X}_1^\dagger = k_2 \underset{\sim}{X}_2^\dagger + \ldots + k_p \underset{\sim}{X}_p^\dagger \end{cases} \quad (3.26)$$

$$\underset{\sim}{X}_1 = k_2 \underset{\sim}{X}_2 + \ldots + k_p \underset{\sim}{X}_p \quad (3.27)$$

which contradicts the assumption of $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_p$ being linearly independent.

If we define

$$\begin{cases} \underset{\sim}{X}_1^{*\dagger} = \underset{\sim}{X}_1^\dagger - k_2 \underset{\sim}{X}_2^\dagger - \ldots - k_p \underset{\sim}{X}_p^\dagger \\ \underset{\sim}{X}_1^{*\dagger} = \underset{\sim}{X}_1^\dagger - k_2 \underset{\sim}{X}_2^\dagger - \ldots - k_p \underset{\sim}{X}_p^\dagger \end{cases} \quad (3.28)$$

then (3.25) may be expressed as

$$\underset{\sim}{X}_1^{*\dagger} = \phi^* \underset{\sim}{X}_1^{*\dagger} \quad (3.29)$$

which (much like (3.21)) determines $\underset{\sim}{X}_1^*$ up to a multiplicative constant, where

$$\underset{\sim}{X}_1^* = \underset{\sim}{X}_1 - k_2 \underset{\sim}{X}_2 - \ldots - k_p \underset{\sim}{X}_p . \quad (3.30)$$

To see that a second or higher order singularity (i.e. $\operatorname{rank}(\underset{\approx}{\Xi}) < p-1$ for $\phi = \phi^*$) can never occur, let us suppose that a second linear decomposition of $\underset{\sim}{\Xi}_1$ existed:

$$\underset{\sim}{\Xi}_1 = k_2' \underset{\sim}{\Xi}_2 + \ldots + k_p' \underset{\sim}{\Xi}_p \quad (3.31)$$

then

$$\underset{\sim}{X}_1^{**} = \underset{\sim}{X}_1 - k_2' \underset{\sim}{X}_2 - \ldots - k_p' \underset{\sim}{X}_p \quad (3.32)$$

would of course also satisfy (3.29), which implies that $\underset{\sim}{X}_1^*$ and $\underset{\sim}{X}_1^{**}$ are identical up to a multiplicative constant. But since $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_p$ are linearly independent this can only happen if $k_2 = k_2', \ldots, k_p = k_p'$ .

Finally defining

$$\underset{\approx}{\Xi}^* = [\underset{\sim}{\Xi}_1^* \ \underset{\sim}{\Xi}_2 \ \cdots \ \underset{\sim}{\Xi}_p] \quad (3.33)$$

with $\underset{\sim}{\Xi}_1^* = \underset{\sim}{X}_1^{*\dagger} - \phi \underset{\sim}{X}_1^{*\dagger}$

we have in a neighbourhood around $\phi = \phi^*$, that

$$SS(\hat{\underset{\sim}{\theta}},\phi) = \underset{\sim}{z}'(I - \underset{\approx}{\Xi}(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}')\underset{\sim}{z}$$
$$= \underset{\sim}{z}'(I - \underset{\approx}{\Xi}^{*}(\underset{\approx}{\Xi}^{*\prime}\underset{\approx}{\Xi}^{*})^{-1}\underset{\approx}{\Xi}^{*\prime})\underset{\sim}{z} \qquad (3.34)$$

hence the general singularity of $\underset{\approx}{\Xi}$ at $\phi = \phi^{*}$ is now expressed in terms of a vanishing column $\underset{\sim}{\Xi}_1^{*}$ in $\underset{\approx}{\Xi}^{*}$, i.e. as a simple singularity considered above.

3.2.4 Singularity in polynomial regression.

A situation of particular interest arise, when a mean $\theta_1$ creates a singularity when $\phi = 1$, but the presence of a deterministic linear trend in the model also makes $\underset{\approx}{\Xi}^{1} = [\underset{\sim}{1}\ \underset{\approx}{\Xi}_{(1)}]$ singular when $\phi = 1$, so that the last step in (3.16) does not go through. Still in a neighbourhood around $\phi = 1$ we have

$$SS(\hat{\underset{\sim}{\theta}},\phi) = \underset{\sim}{z}'(I - \underset{\approx}{\Xi}(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}')\underset{\sim}{z}$$
$$= \underset{\sim}{z}'(I - \underset{\approx}{\Xi}^{1}(\underset{\approx}{\Xi}^{1\prime}\underset{\approx}{\Xi}^{1})^{-1}\underset{\approx}{\Xi}^{1\prime})\underset{\sim}{z}\ . \qquad (3.35)$$

If without loss of generality we say that $\underset{\sim}{X}_2 = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$, then

$$\underset{\sim}{\Xi}_2 = \begin{bmatrix} 1 + (1-\phi) \\ 1 + 2(1-\phi) \\ \vdots \\ 1 + (n-1)(1-\phi) \end{bmatrix} \qquad (3.36)$$

now defining

$$\underset{\approx}{\Xi}^{*} = [\underset{\sim}{\Xi}_2^{*}\ \underset{\sim}{1}\ \underset{\sim}{\Xi}_3 \cdots \underset{\sim}{\Xi}_p] \qquad (3.37)$$

where

$$\underset{\sim}{\Xi}_2^{*} = \underset{\sim}{\Xi}_2 - \underset{\sim}{1} = \begin{bmatrix} (1-\phi) \\ 2(1-\phi) \\ \vdots \\ (n-1)(1-\phi) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n-1 \end{bmatrix} \qquad (3.38)$$

it is recognized, that letting $\theta_2$ retain its interpretation as a



deterministic trend at the point $\phi = 1$, i.e. manifesting itself as a mean for the differences, the "replacement parameter" $\theta_1^{*}$ for $\theta_1$ has the interpretation of a deterministic linear trend for the differences.

Higher order polynomial trends may obviously be handled similarly.

3.2.5 Summary of findings.

Lemma 1

Let $\underset{\approx}{\Xi}_{(n-1)\times p} = \underset{\approx}{C}_{(n-1)\times n}\ \underset{\approx}{X}_{n\times p} = \begin{bmatrix} -\phi & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -\phi & 1 \end{bmatrix} \underset{\approx}{X}$

and let $\underset{\approx}{X}$ be any real matrix of full column rank, p. Then $\underset{\approx}{\Xi}$ has also rank p except possibly at distinct points $\phi = \phi^{*}$, where $\text{rank}(\underset{\approx}{\Xi}) = p-1$.

Lemma 2

For $\phi \neq \phi^{*}$, denote by $SS(\hat{\underset{\sim}{\theta}},\phi)$ the residual sum of squares from fitting by least squares the (transformed) linear model $\underset{\sim}{z} = \underset{\approx}{\Xi}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}$, where $\underset{\sim}{z} = \underset{\approx}{C}\underset{\sim}{y}$, $\underset{\sim}{y}$ is any n-dimensional vector (of "original" observations) and $\underset{\sim}{\varepsilon}$ is the vector of random disturbances. Then $SS(\hat{\underset{\sim}{\theta}},\phi)$ has a limiting value as $\phi \to \phi^{*}$, and it becomes a continuous function of $\phi \in R$ if, by definition, we set $SS(\hat{\underset{\sim}{\theta}},\phi^{*}) = \lim_{\phi\to\phi^{*}} SS(\hat{\underset{\sim}{\theta}},\phi)$.

Corollary

If $\underset{\sim}{\Xi}_j = \underset{\sim}{0}$ for $\phi = \phi^{*}$ then $SS(\hat{\underset{\sim}{\theta}},\phi^{*})$ may be computed as the residual sum of squares from fitting by least squares the linear model $\underset{\sim}{z} = \underset{\approx}{\Xi}^{*}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}$ provided $\text{rank}(\underset{\approx}{\Xi}^{*}) = p$, where

$$\underset{\approx}{\Xi}^{*} = [\underset{\sim}{\Xi}_1 \cdots \underset{\sim}{\Xi}_{j-1}\underset{\sim}{X}_{j+}\underset{\sim}{\Xi}_{j+1} \cdots \underset{\sim}{\Xi}_p]$$

and $\underset{\sim}{X}_{j+}$ is the j-th column of $\underset{\approx}{X}$ having deleted the n-th row. The interpretations of the $\theta$'s are unchanged when $\phi = \phi^{*}$ except for $\theta_j$

which at that point assumes a different role.

In particular if $\underset{\sim}{X}_j = \underset{\sim}{1}$, i.e. the linear model involves a mean, then $\phi = 1$ is a singularity point. If further the model does not involve a deterministic trend, then at this point $\theta_1$ has the interpretation of a mean for the differenced observations, $y_2-y_1$, $y_3-y_2$, ..., $y_n-y_{n-1}$. If the model includes polynomial trends up to and including the order $q$, then $\theta_j$ has the interpretation of a $q$-th order polynomial trend for the differences.

## 3.3 Inference about $\phi$ based on maximized likelihood.

The likelihood function corresponding to the model (3.1) and (3.2) may be written as (Equation (2.29) in Chapter 2):

$$\ell(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = \sigma^{-n}\exp(-\tfrac{1}{2}\sigma^{-2}((\underset{\sim}{z}-\underset{\approx}{\Xi}\underset{\sim}{\theta})'(\underset{\sim}{z}-\underset{\approx}{\Xi}\underset{\sim}{\theta}) + (e_1-M)^2) \qquad (3.39)$$

where $M$ is a starting parameter for the noise:

$$e_1 = M + \varepsilon_1 \quad . \qquad (3.40)$$

If it is known a priori that the noise is stationary and reversible, then $e_1 = (1-\phi^2)^{-1/2}\varepsilon_1$ (see for example Anderson [1954]); but since such knowledge will seldom be available, $M$ is included to hedge against the possibility that the first observation might have an untypical error, and more importantly to broaden the noise structure to cover explosive as well as stationary AR-1 processes (see Section 2.1 of Chapter 2).

Maximizing the likelihood (3.39) with respect to $\{M,\underset{\sim}{\theta},\phi,\sigma\}$ it is found that

$$\max_{\{M,\underset{\sim}{\theta},\phi,\sigma\}} \ell(M,\underset{\sim}{\theta},\phi,\sigma|\underset{\sim}{y}) = \left(\frac{SS(\hat{\underset{\sim}{\theta}},\hat{\phi})}{n}\right)^{-n/2} e^{-n/2} \qquad (3.41)$$

where

$$SS(\hat{\underset{\sim}{\theta}},\hat{\phi}) = \min_{\phi} SS(\hat{\underset{\sim}{\theta}},\phi) \qquad (3.42)$$



It is noted, that the maximum likelihood estimate (MLE) of $\phi$, is the value $\hat{\phi}$ which minimizes the residual sum of squares left by an ordinary least squares fit of the transformed linear model (3.8).

A test for a hypothesis of $\phi = \phi^\circ$ may be constructed from the likelihood ratio

$$\lambda_{\phi^\circ} = \frac{\max_{\{M,\underset{\sim}{\theta},\phi=\phi^\circ,\sigma\}} (\ell)}{\max_{\{M,\underset{\sim}{\theta},\phi,\sigma\}} (\ell)} = \frac{\ell(\hat{M}_o,\hat{\underset{\sim}{\theta}}_o,\phi=\phi^\circ,\hat{\sigma}_o|\underset{\sim}{y})}{\ell(\hat{M},\hat{\underset{\sim}{\theta}},\hat{\phi},\hat{\sigma}|\underset{\sim}{y})} = \left(\frac{SS(\hat{\underset{\sim}{\theta}}_o,\phi=\phi^\circ)}{SS(\hat{\underset{\sim}{\theta}},\hat{\phi})}\right)^{-n/2} \qquad (3.43)$$

by considering the well known large sample $\chi^2$ distribution approximation of the likelihood ratio statistic

$$\Lambda_{\phi^\circ} = -2\ell n\, \lambda_{\phi^\circ} \overset{\sim}{\sim} \chi^2_1 \quad . \qquad (3.44)$$

This approximation is justified as the asymptotic distribution of $\Lambda_{\phi^\circ}$ (under $\phi = \phi^\circ$), when the maximum likelihood estimators are asymptotically normally distributed (see e.g. Rao [1965] or Wilks [1962]). Usually the asymptotic normality of MLE is established for independent observations, but Whittle [1953] has extended the limiting properties of MLE to cover stationary time series. More specifically Hildreth [1969] showed, that the MLE of the parameters of a linear model with stationary first order autoregressive errors have an asymptotic multivariate normal distribution with mean vector equal to the true parameter values, and the estimators are asymptotically efficient. Hence the $\chi^2$ approximation for the distribution of $\Lambda_{\phi^\circ}$ holds for $-1 < \phi^\circ < 1$.

For $|\phi^\circ| > 1$, i.e. for explosive series, it does not make sense to talk about asymptotic results, and in fact it is demonstrated in Section 3.6, that (3.44) does not hold in this case. Exceptions must also be made for singularity points $\phi^\circ = \phi^*$. For although $SS(\hat{\underset{\sim}{\theta}},\phi)$ can be made continuous at these points, it was seen in Section 3.2 that

the continuity required an artificial inclusion of a substitution parameter, which is not part of the original model, hence a test for $\phi^\circ = \phi^*$ using (3.44) would be conditional on an implicit misspecification of the model at this very point.

In particular (3.44) suggests an approximate likelihood ratio test for $\phi = 0$ based on

$$\Lambda_0 = -2\ell n\, \lambda_0 = n\, \ell n\, SS(\hat{\underset{\sim}{\theta}}_0, \phi = 0) - n\, \ell n\, SS(\hat{\underset{\sim}{\theta}}, \hat{\phi}) \sim \chi^2_1 \qquad (3.45)$$

(How well this approximation works in moderate sample sizes is studied by simulation in Section 3.6).

In practice rather than merely testing whether $\Lambda_0$ takes a significantly large value it is usually most informative to plot

$$LMAX(\phi) = -n\, \ell n\, SS(\hat{\underset{\sim}{\theta}}, \phi) \qquad (3.46)$$

over a suitable range of $\phi$. The mode of the $LMAX(\phi)$-curve marks $\hat{\phi}$ and an approximate $100(1-\alpha)\%$ confidence interval for $\phi$ is obtained from

$$LMAX(\hat{\phi}) - LMAX(\phi) < \chi^2_1(\alpha) \; . \qquad (3.47)$$

In other words the plausibility of any value of $\phi$ may be approximately assessed from such a plot, and it becomes visually clear whether the data set at hand contains much information about $\phi$ or only relatively little.

Ordinarily $\phi$ is a nuisance parameter in relation to the linear parameters $\underset{\sim}{\theta}$, and inferences about $\phi$ may only be of interest as a stepping stone to the further analysis concerning $\underset{\sim}{\theta}$. Unless strong reasons exist for believing that $\phi = \phi^\circ (=0)$, the occurrence of $\phi^\circ$ inside a certain confidence internal supplies little justification for outright assuming that $\phi = \phi^\circ$. From a likelihood point of view it would seem more natural to base inferences about $\underset{\sim}{\theta}$ on $\phi = \hat{\phi}$, and



hence to estimate the independence inducing transformation $\underset{\approx}{c}$ simultaneously with $\underset{\sim}{\theta}$.

## 3.4 Bayesian inference about $\phi$.

From the Bayesian point of view all that can be said about $\phi$ in light of the data, is embodied in the marginal posterior distribution of $\phi$, $p(\phi|\underset{\sim}{y})$. As was shown in Chapter 2, the posterior density function of $\phi$ resulting from multiplying the likelihood function $\ell$ in (3.39) by a prior $p(M,\underset{\sim}{\theta},\phi,\sigma)$ which is approximately noninformative for $\{M,\underset{\sim}{\theta},\sigma\}$, and integrating out $\{M,\underset{\sim}{\theta},\sigma\}$ is

$$p(\phi|\underset{\sim}{y}) \propto p(\phi)(SS(\hat{\underset{\sim}{\theta}},\phi))^{-\frac{n-p-1}{2}} \qquad (3.48)$$

where $p(\phi)$ is the marginal prior for $\phi$.

If $p(\phi)$ is chosen uniform, then

$$p^e_u(\phi|\underset{\sim}{y}) \propto SS(\hat{\underset{\sim}{\theta}},\phi)^{-\frac{n-p-1}{2}} \qquad (3.49)$$

where "$_u$" stands for "uniform" and "$^e$" stands for "exact". Obviously $p^e_u(\phi|\underset{\sim}{y})$ has its mode at the least squares value $\phi = \hat{\phi}$; and this distribution may be approximated closely by a t-distribution with $\nu = n-p-2$ degrees of freedom provided $SS(\hat{\underset{\sim}{\theta}},\phi)$ is approximated near $\hat{\phi}$ by

$$Q_t(\phi) = b_1 + b_2\phi + b_3\phi^2 \qquad (3.50)$$

then relating to the t-form

$$p^e_u(\phi|\underset{\sim}{y}) \doteq \left(1 + \frac{(\phi-\hat{\phi})^2}{\nu s_t^2}\right)^{-\frac{\nu+1}{2}} \propto p_u(\phi|\underset{\sim}{y}) \qquad (3.50)$$

we have that

$$p_u(\phi|\underset{\sim}{y}) \sim t(\hat{\phi}, s_t, \nu)$$

$$\text{with } \hat{\phi} = -\frac{b_2}{2b_3} \qquad (3.51)$$

$$s_t = \sqrt{\frac{4b_1b_3 - b_2^2}{4b_3^2\nu}}$$

and $\nu = n-p-2$ .

Alternatively in the vicinity of $\hat{\phi}$ , an approximately noninformative prior for $\phi$ is (see Section 2.5 of Chapter 2)

$$p_n(\phi) \propto ((n-1) + (n-2)\phi^2 + \ldots + \phi^{2(n-2)})^{1/2} \qquad (3.52)$$

where "$_n$" stands for "noninformative" (but also serves to remind, that this prior is a function of $n$). However a direct multiplication by $p_n(\phi)$ makes the right hand side of (3.48) diverge as $\phi$ becomes large. As discussed in Section 2.5 of Chapter 2, the argument leading to this approximately noninformative prior is a local one, applicable in the vicinity of $\hat{\phi}$ only. Here the effect on the posterior of the noninformative prior as compared to a uniform prior, is to shift its mode moderately away from the origin, along with an increase of the spread. These modifications may be accomplished within the t-form by fitting the polynomial

$$Q_\tau = \beta_1 + \beta_2\phi + \beta_3\phi^2 \qquad (3.53)$$

through the three points $Q_\tau(\phi'-\Delta\phi)$, $Q_\tau(\phi')$ and $Q_\tau(\phi'+\Delta\phi)$, calculated from

$$Q_\tau(\phi'\pm\Delta\phi) = q(\phi'\pm\Delta\phi) \cdot Q_t(\phi'\pm\Delta\phi) \qquad (3.54)$$

where $\phi'$ is reasonably close to $\hat{\phi}$ and $\Delta\phi$ is about or less than $s_t$; and where

$$q(\phi) \propto p_n(\phi)^{-\frac{2}{n-p-1}} . \qquad (3.55)$$

The modified quadratic curve (3.53) determines

$$p_n(\phi|\underset{\sim}{y}) \sim t(\hat{\phi}, s_\tau, \nu) \qquad (3.56)$$

where $\hat{\phi}$ , $s_\tau$ and $\nu$ are as in (3.51) except the b's are replaced by $\beta$'s.

The Bayesian parallel to a significance test at the $\alpha$ level of a hypothesis $\phi = \phi^\circ$, consists of determining whether $\phi^\circ$ is or is not



included in the $100(1-\alpha)\%$ highest posterior density (HPD) interval for $\phi$ (see for example Box and Tiao [1972], pp. 122-127). This may now conveniently be checked by referring the quantities

$$t_{\phi^\circ} = \frac{\phi^\circ - \hat{\phi}}{s_t} \qquad (3.57)$$

or

$$\tau_{\phi^\circ} = \frac{\phi^\circ - \hat{\phi}}{s_\tau} \qquad (3.58)$$

to a $t(0,1,n-p-2)$ distribution depending on which prior is adopted. In this treatment the plausibility of any hypothesized value $-\infty < \phi^\circ < \infty$ may be "tested", except singularity points $\phi^\circ = \phi^*$ . In particular this excludes (as before) $\phi^\circ = 1$ for models involving a mean. This is unfortunate since $\phi^\circ = 1$ implying that the noise is a random walk may be of special interest (this question shall be returned to in Chapter 4).

From a Bayesian viewpoint in making inferences about the linear parameters $\underset{\sim}{\theta}$ , the whole posterior distribution $p(\phi|\underset{\sim}{y})$ should be used; and we have seen that utilizing the familiar t form, $p(\phi|\underset{\sim}{y})$ may be approximated with a modest computational effort.

*Specifically*

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int_R p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y})p(\phi|\underset{\sim}{y})d\phi \qquad (3.59)$$

where

$$p(\underset{\sim}{\theta}|\phi,\underset{\sim}{y}) \sim t_p(\hat{\underset{\sim}{\theta}}, s^2(\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}, \nu)$$

with $$\hat{\underset{\sim}{\theta}} = (\underset{\approx}{\Xi}'\underset{\approx}{\Xi})^{-1}\underset{\approx}{\Xi}'\underset{\sim}{z} \qquad (3.60)$$

$$\nu s^2 = SS(\hat{\underset{\sim}{\theta}}, \phi)$$

and $$\nu = n-p-1 .$$

When the posterior (3.59) may be approximated by

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) \doteq p(\underset{\sim}{\theta}|\hat{\phi},\underset{\sim}{y}) \qquad (3.61)$$

the Bayesian inference about $\underset{\sim}{\theta}$ exactly parallels the likelihood solution suggested in the preceding Section 3.3.

## 3.5 Examples.

In this section we shall reexamine three data sets used as examples in previously published papers.

The first set is one generated by Zellner and Tiao [1964], and the "true" process does not include a mean. The second set is the textile data examined in the paper by Theil and Nagar [1961]. The third set is the "spirits series" considered by Durbin and Watson [1951], and re-examined by Theil and Nagar [1961].

### 3.5.1 The data generated by Zellner and Tiao.

Zellner and Tiao [1964] generated and analyzed the data in Table 3.I, using the model:

$$y_i = \theta x_i + e_i \quad \text{with} \quad \theta = 3. \tag{3.62}$$

$$\begin{cases} e_1 = M + \epsilon_1 & \text{with} \quad M = .5 \\ e_i = \phi e_{i-1} + \epsilon_i & \text{with} \quad \phi = .5 \end{cases} \tag{3.63}$$

where the $\epsilon_i$'s are i.i.d. N(0,1).

Probing for serial correlation along the lines of Section 3.3, LMAX($\phi$) (3.46) is plotted in Figure 3.1a. LMAX($\phi$) has its maximum at $\hat{\phi} = .36$, and the upper horizontal bars in the plot marks the level of LMAX($\hat{\phi}$). The lower set of bars are drawn in a distance of $3.84 = \chi^2_1(.05)$ from LMAX($\hat{\phi}$), hence an approximate 95% confidence interval for $\phi$ is read off as $-.14 < \phi < .88$. The evidence of positive serial correlation is not very strong, in particular $\Lambda_0$ is not significant at the 5% level. (The points LMAX($\phi$=0) and LMAX($\phi$=1) are indicated by vertical bars on the LMAX($\phi$) graph).



TABLE 3.1, Zellner and Tiao's data.

| i | $y_i$ | $x_i$ |
|---|---|---|
| 1 | 12.649 | 3.9 |
| 2 | 18.794 | 6.0 |
| 3 | 12.198 | 4.2 |
| 4 | 14.372 | 5.2 |
| 5 | 13.909 | 4.7 |
| 6 | 14.556 | 5.1 |
| 7 | 14.700 | 4.5 |
| 8 | 18.281 | 6.0 |
| 9 | 13.890 | 3.9 |
| 10 | 10.318 | 4.1 |
| 11 | 5.473 | 2.2 |
| 12 | 4.044 | 1.7 |
| 13 | 6.361 | 2.7 |
| 14 | 7.036 | 3.3 |
| 15 | 13.368 | 4.8 |

Figure 3.1b gives the Bayesian picture of what the data has to tell about $\phi$. The solid curve is $p_n(\phi|\underset{\sim}{y})$ with mode at $\hat{\phi} = .39$, and its 95% HPD interval for $\phi$ is $-.25 < \phi < 1.01$. The two broken curves in Figure 3.1b, $p_u^e(\phi|\underset{\sim}{y})$ and $p_u(\phi|\underset{\sim}{y})$, agree almost perfectly; also it makes little difference if inference about $\phi$ were based on those distributions rather than on $p_n(\phi|\underset{\sim}{y})$.

Turning to the DW-test it is perhaps somewhat surprising to find $d = 2.002$, i.e. $\check{\phi} = -.001$. Of course in order to compute $d$, it is necessary to artifically include a mean in the linear model, i.e. relate the data to the model

$$y_i = \theta_1 + \theta_2 X_i + e_i \tag{3.64}$$

where $e_i$ is as in (3.63).

It is interesting to redraw LMAX($\phi$) and $p(\phi|\underset{\sim}{y})$ for this modified model (3.64); this is done in Figure 3.2. It discloses, that incorporating a mean into the model, destroys all traces of positive serial correlation which are imprinted in the data. Further insight into the problems created by model misspecification may be gained by considering what inferences can be drawn about the linear parameters $\theta_1$ and $\theta_2$ of (3.64). Figure 3.3a shows posterior distributions of the mean $\theta_1$. The graph $p(\theta_1|\phi=0,\underset{\sim}{y})$ is also the confidence distribution of $\theta_1$ from a sampling theory point of view, since it is identical to the distribution of the statistic $\hat{\theta}_1$ conditional on $\phi = 0$. Since in this framework according to the procedure recommended by Durbin and Watson the value $\phi = 0$ would be accepted, then $\theta_1$ would subsequently be found significantly different from zero at the $\alpha = .005$ level (two sided), and $\theta_1$ could be mistakenly retained in the model. The distribution $p(\theta_1|\phi = \hat{\phi},\underset{\sim}{y})$ can also be interpreted as a confidence distribution for $\theta_1$ given that $\phi = \hat{\phi} = .36$. In this case the significance of $\theta_1$ has dropped to the $\alpha = .03$ level (two sided). In contrast a Bayesian analysis does not discount the possibility of $\theta_1$ being zero, as this point lies well within the 90% HPD region of $p_n(\theta_1|\underset{\sim}{y})$ (and of $p_u(\theta_1|\underset{\sim}{y})$).

It is of further interest to see what impact the inclusion of the mean, $\theta_1$, has on making inferences about the regression coefficient which appears as $\theta$ in the model (3.62) and as $\theta_2$ in the model (3.64), and which is most likely to be the parameter of primary interest. Figures 3.3b and 3.3c show posterior distributions for $\theta_2$ and $\theta$ respectively. It is seen, that not only has the inclusion of $\theta_1$ the effect of diluting the information in the data about $\theta$ ($\theta_2$), but $p(\theta_2|\phi = 0,\underset{\sim}{y})$ in Figure 3.3b actually throws doubt on the plausibility



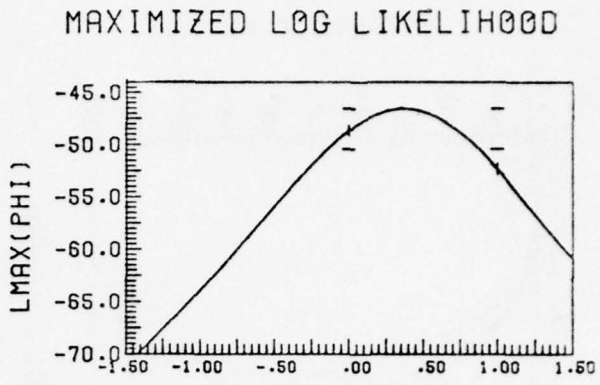



FIGURE 3.1a

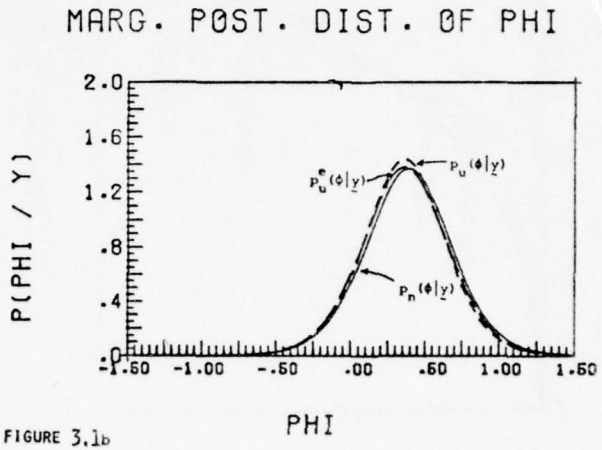



FIGURE 3.1b

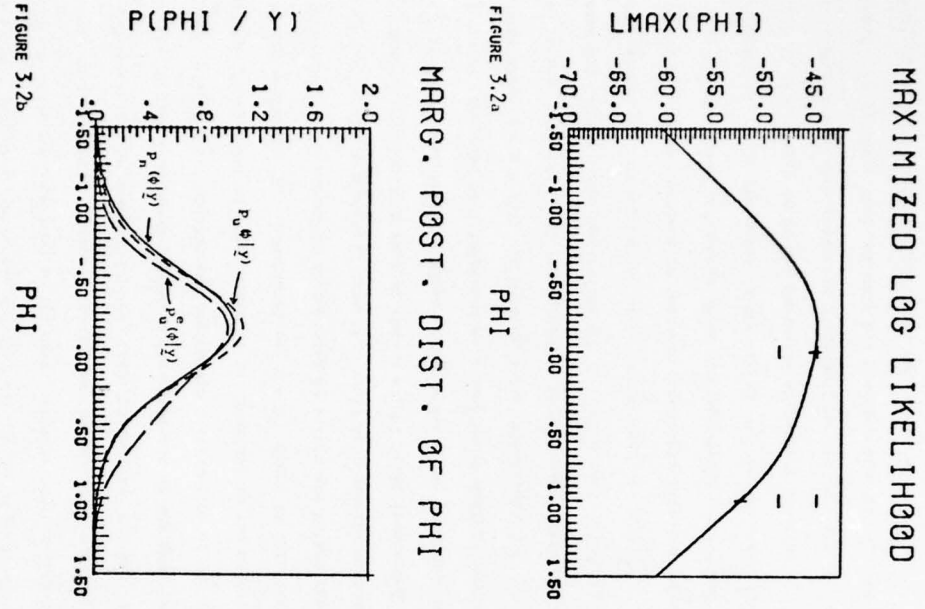



FIGURE 3.2a




FIGURE 3.2b



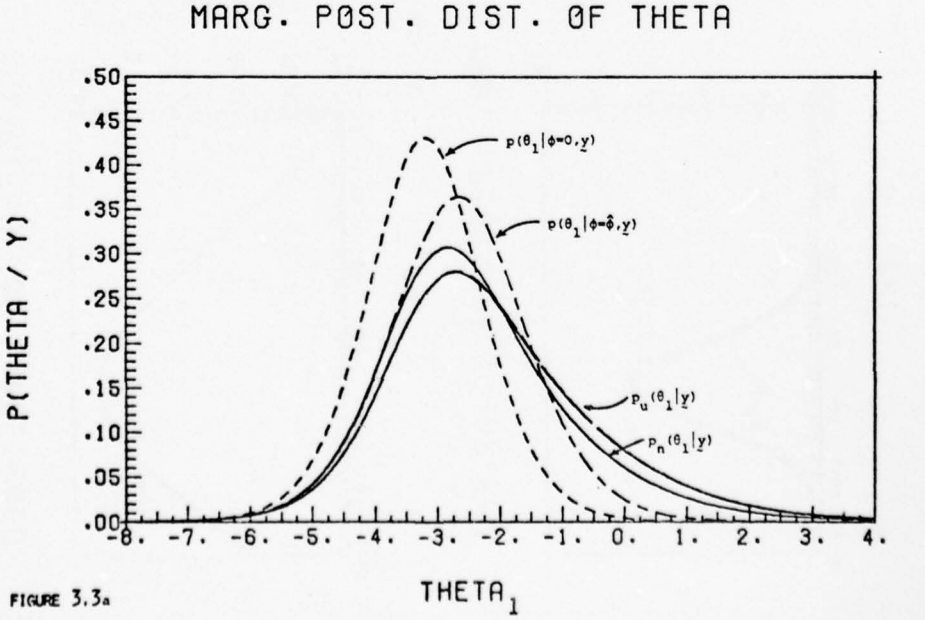



FIGURE 3.3a

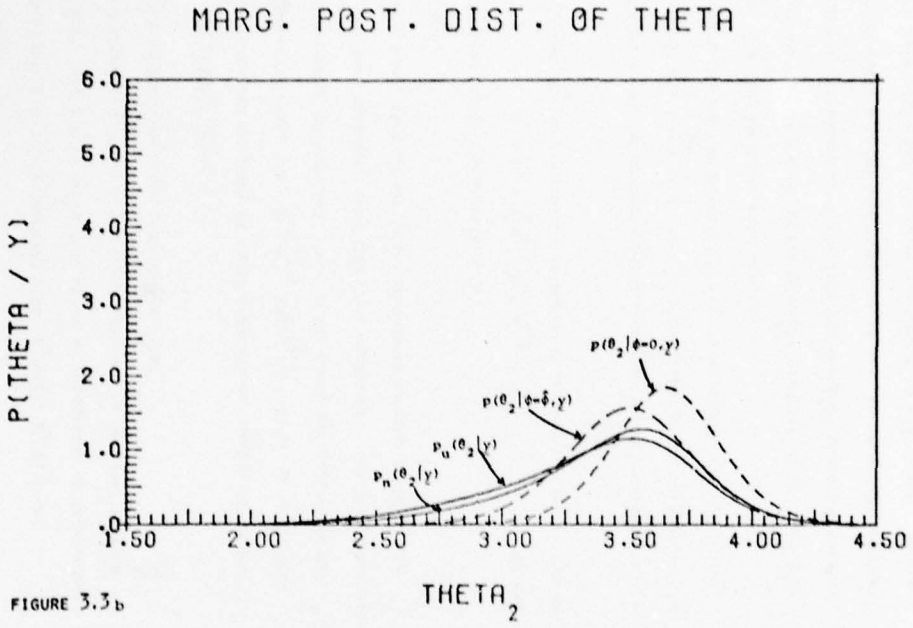



FIGURE 3.3b



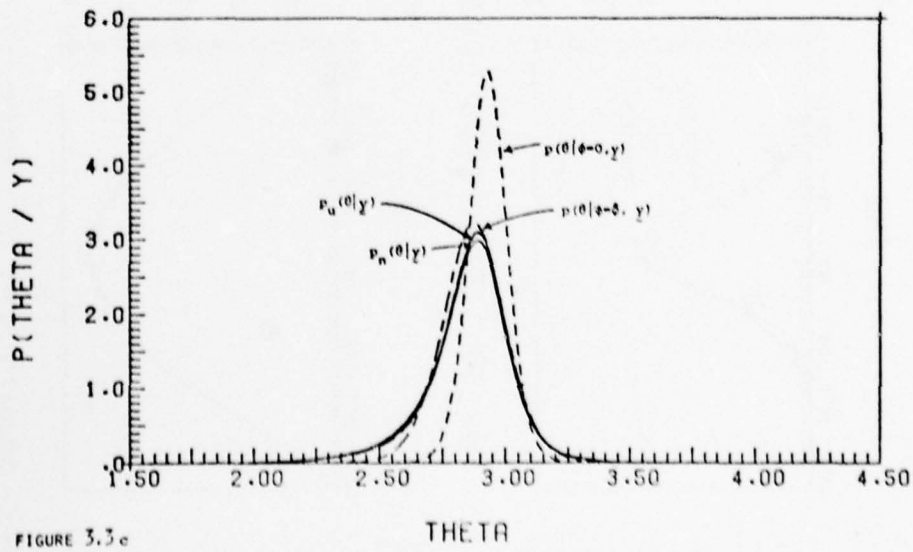



FIGURE 3.3c

of the true value $\theta_2 = 3$. On the other hand $p(\theta_2|\underset{\sim}{y})$ covers $\theta_2 = 3$. nicely, while this value is just included in the 95% HPD region of $p(\theta_2|\phi = \hat{\phi},\underset{\sim}{y})$.

In Figure 3.3c it is observed that $p_n(\theta|\underset{\sim}{y})$, $p_u(\theta|\underset{\sim}{y})$ and $p(\theta|\phi = \hat{\phi},\underset{\sim}{y})$ are almost coincidal, while an assumption of independence expressed through $p(\theta|\phi = 0,\underset{\sim}{y})$ would lead one to believe, that $\theta$ is known with greater precision than is justified.

3.5.2 The textile data.

This data set consists of three simultaneous series of length 17 (yearly observations), viz. $y_i$, $x_{i,1}$ and $x_{i,2}$, where $y_i$ = log (per capita consumption of textile), $x_{i,1}$ = log (real per capita income), and $x_{i,2}$ = log (deflated price index for clothing). The data shall not be relisted here, they appear along with their sources in Theil and Nagar [1961].

The model under consideration is

$$y_i = \theta_1 + \theta_2 x_{i,1} + \theta_3 x_{i,2} + e_i \qquad (3.65)$$

where $\theta_2$ and $\theta_3$ are sometimes referred to as elasticities in economics.

Fitting (3.65) by ordinary least squares, the DW-statistic is calculated to be $d = 1.926$ ($\tilde{\phi} = .04$), so that data set has been used as an example where the null hypothesis $\phi = 0$ is accepted, and inference about the $\theta$'s may be drawn accordingly.

Studying Figure 3.4 as in the previous example, it is found both from maximized likelihood plot and from the marginal posterior density of $\phi$, that while a prior hypothesis of $\phi = 0$ is certainly not discounted, a very wide range of other hypotheses $\phi = \phi^*$ cannot be dismissed either. Specifically Figure 3.4a yields the approximate 95%



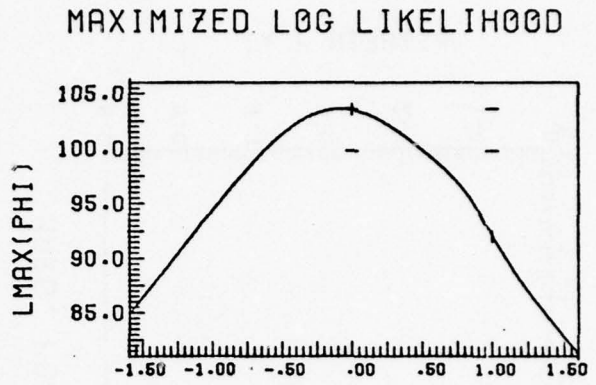



FIGURE 3.4a

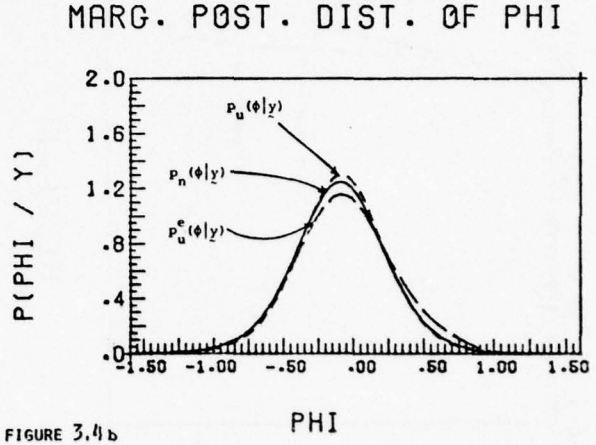



FIGURE 3.4b

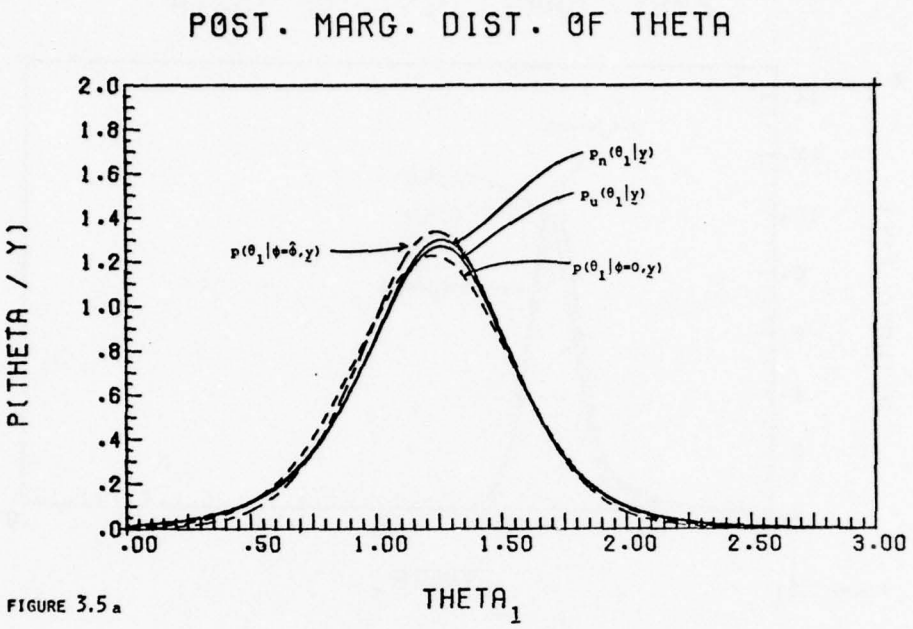



FIGURE 3.5a



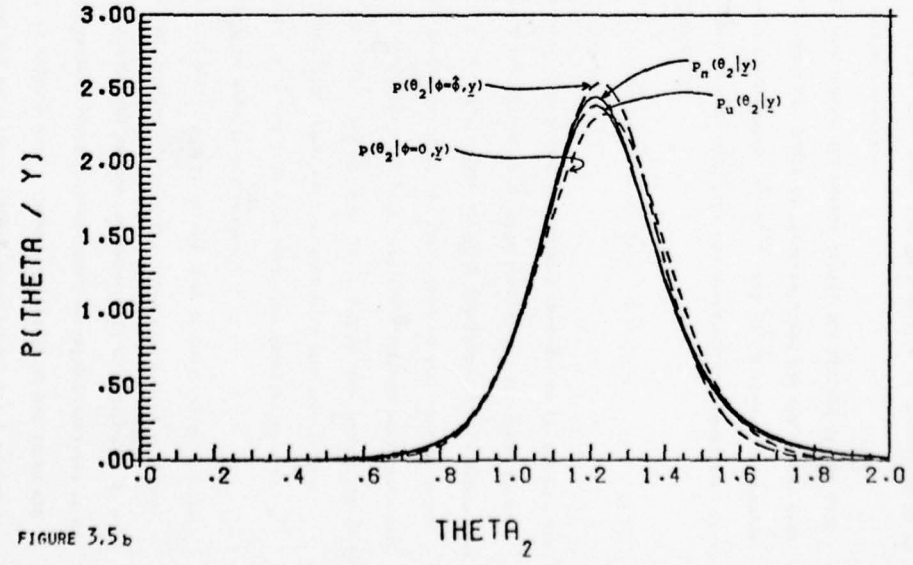



FIGURE 3.5b

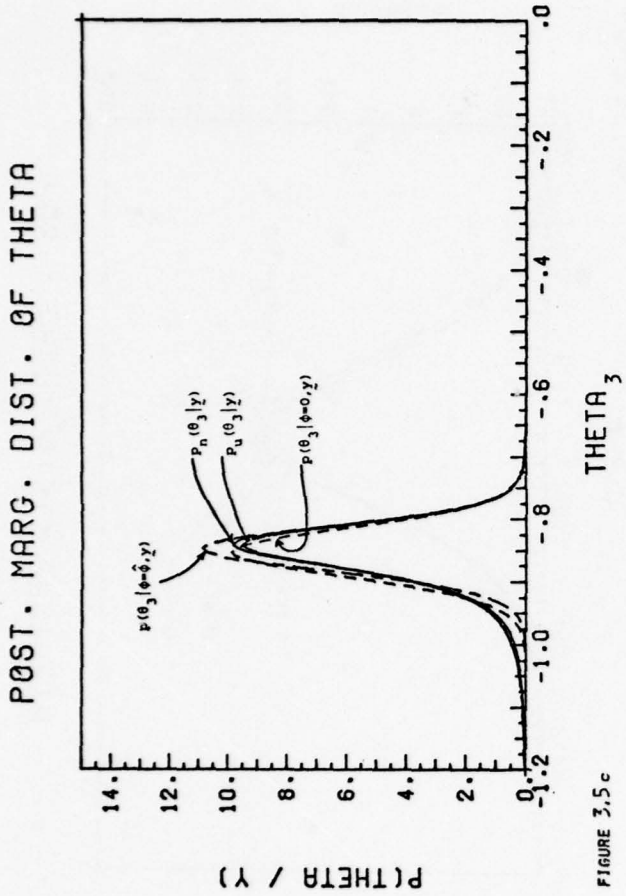



FIGURE 3.5c



confidence interval $-.63 < \phi < .51$, with $\hat{\phi} = -.09$; and Figure 3.4b show a 95% HPD region of $-.78 < \phi < .59$, with $\overset{*}{\phi} = -.10$.

The question may be raised, whether there is any prior reason to suspect that $\phi$ is exactly zero. Causes can easily be put forth why observations of this kind would be positively serially correlated (e.g. gradual change in purchasing habits) or negatively correlated (e.g. a cold winter may spur purchasing, with the after effect of a sluggish market the following year); but it is not easy to rationalize why the observations should be exactly independent.

Figures 3.5a, 3.5b and 3.5c show posterior distributions of $\theta_1$, $\theta_2$ and $\theta_3$ respectively. For each $\theta_j$ parameter the four curves $p_n(\theta_j|\underset{\sim}{y})$, $p_u(\theta_j|\underset{\sim}{y})$, $p(\theta_j|\ \phi=\hat{\phi},\underset{\sim}{y})$ and $p(\theta_j|\ \phi=0,\underset{\sim}{y})$ are almost identical. It is therefore a fact, that for this particular example where the mean of the posterior distribution of $\phi$ lies close to zero correct inferences about the $\theta_j$'s result from assuming independence. It would seem however, that this has little to do with testing and accepting an unlikely hypothesis of $\phi = 0$, but is rather a consequence of the circumstance that for this data

$$p(\theta_j|\phi=0,\underset{\sim}{y}) \doteq p(\theta_j|\phi=\hat{\phi},\underset{\sim}{y}) \doteq p(\theta_j|\underset{\sim}{y}) \qquad j = 1,2,3 \tag{3.66}$$

3.5.3 The spirits data.

This data set is very much like the textile data, consisting of three simultaneous time series $y_i$, $x_{i,1}$ and $x_{i,2}$, defined as before except "spirits" takes the place of "textile", and now the series have 69 observations (see Durbin and Watson [1951] for listing of the data and reference to their source).

Fitting (3.65) to this data, the DW-statistic is calculated to be $d = .249$ ($\hat{\phi} = .875$), which is significant at the 2% level (two sided).

The maximized likelihood curve in Figure 3.6a shows that $\hat{\phi} = 1.015$. An approximate 95% confidence interval canstructed as described in Section 3.5.1 is $.98 < \phi < 1.05$; however this interval cannot be relied on, since the $\chi^2$ approximation (3.44) only finds theoretical justification for $-1 < \phi < 1$. The Bayesian probe for serial correlation, Figure 3.6b, determines from $p_n(\phi|y)$ a 95% HPD interval of $.99 < \phi < 1.06$, with $\overset{*}{\phi} = 1.025$.

Uniform agreement prevails that the observations are strongly serially correlated. Theil and Nagar [1961] carry out the further analysis concerning $\underset{\sim}{\theta}$ as if $\phi = \check{\phi}$. But clearly $\check{\phi}$ is a very poor estimate of $\phi$. Approaching the problem from a sampling theory point of view it would seem most reasonable to use $\hat{\phi}$ rather than $\check{\phi}$ in the independence inducing transformation (3.9) and then analyse (3.8) accordingly. Because while $\hat{\phi}$ is the maximum likelihood estimate of $\phi$, $\check{\phi}$ is a maximum likelihood estimate conditional on $\underset{\sim}{\theta} = \check{\underset{\sim}{\theta}} = (\underset{\approx}{X}'\underset{\approx}{X})^{-1}\underset{\approx}{X}'\underset{\sim}{y}$ which may be a quite misleading estimate if $\phi$ is not close to zero. In the Bayesian framework when $p(\phi|\underset{\sim}{y})$ is very narrow as here, then the conditional (on $\phi = \overset{*}{\phi}$, say) distribution of $\underset{\sim}{\theta}$ will generally be quite similar to the marginal distribution.

The results of alternative approaches are shown in Figures 3.7a, 3.7b and 3.7c for $\theta_1$, $\theta_2$ and $\theta_3$ respectively. For $\theta_2$ and $\theta_3$ the three curves $p_n(\theta_j|\underset{\sim}{y})$, $p_u(\theta_j|\underset{\sim}{y})$ and $p(\theta_j|\phi=\hat{\phi},\underset{\sim}{y})$ are very much alike, where the latter one also may be interpreted as a confidence distribution for $\theta_j$. An assumption of independence renders a grossly erroneous impression of the linear parameters $\underset{\sim}{\theta}$; but also the substitution of $\phi = \check{\phi}$, while not quite as bad as $\phi = 0$, yields misleading results. And it is not just a matter of overestimating the precision



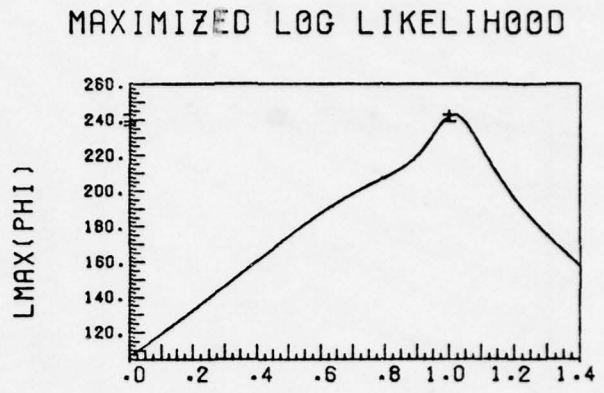



FIGURE 3.6a

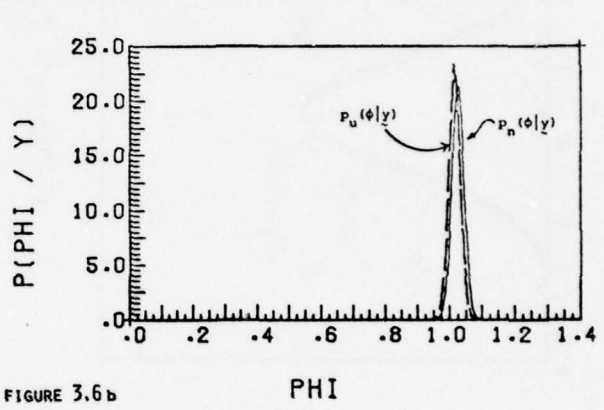



FIGURE 3.6b

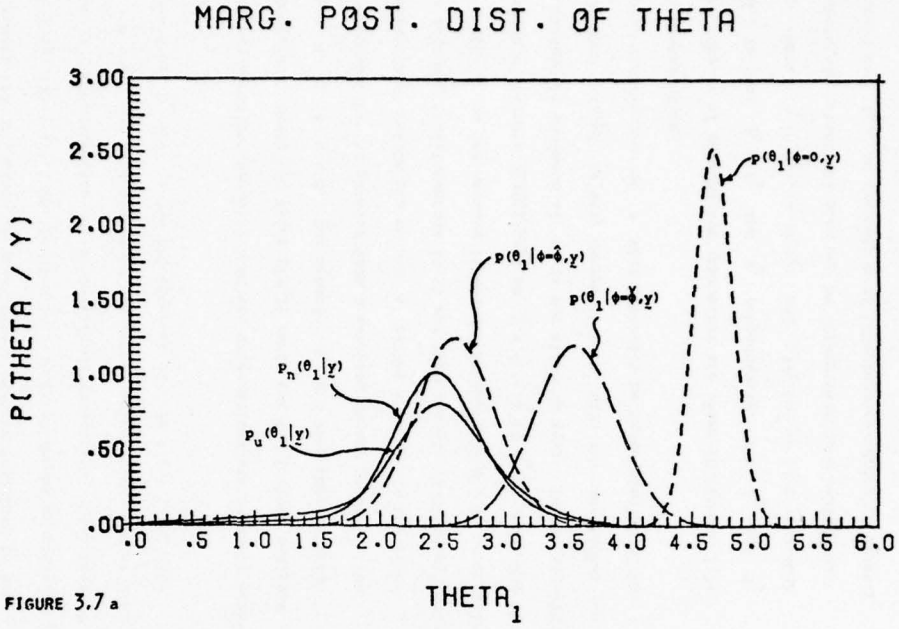



FIGURE 3.7 a



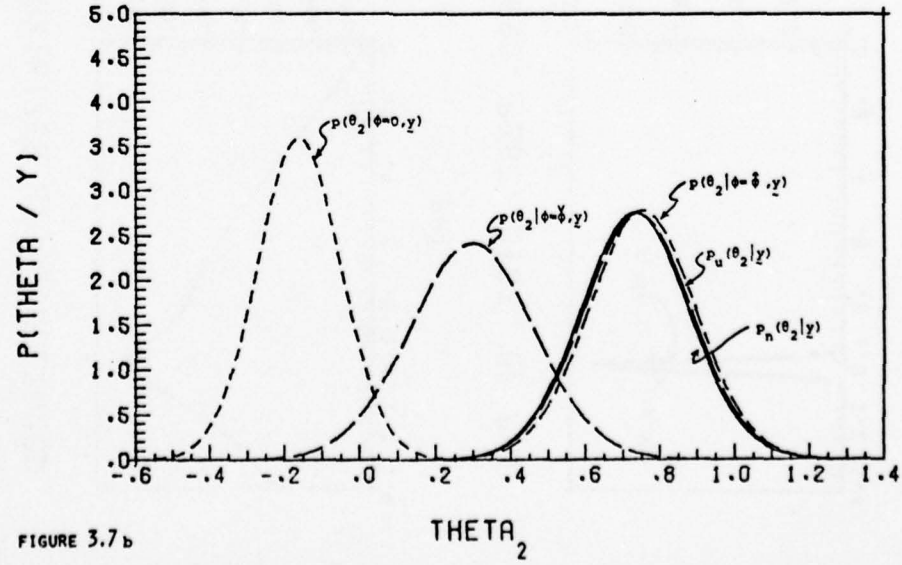



FIGURE 3.7 b

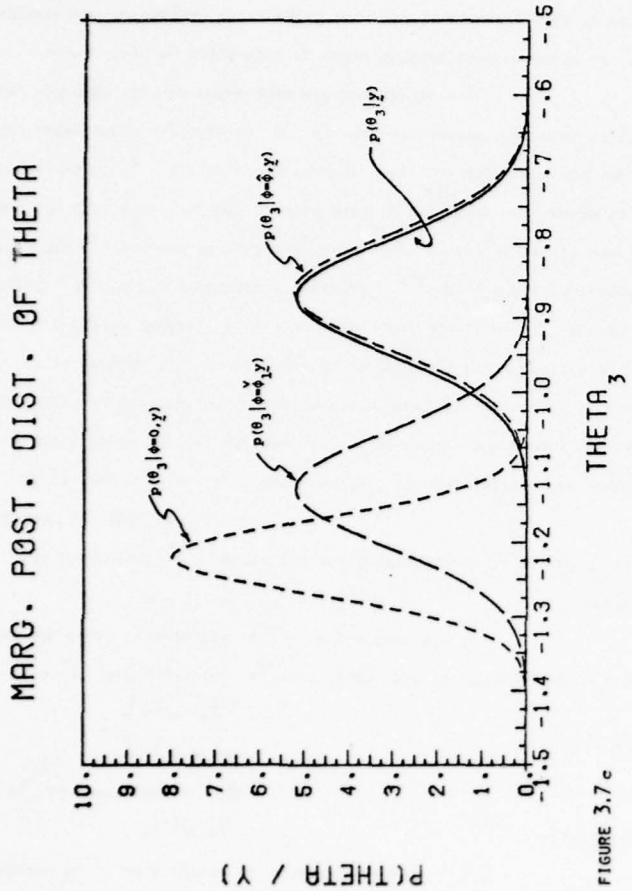



FIGURE 3.7c



with which the $\theta_j$'s are known, but one of location. Specifically the income elasticity is actually much greater than was suggested by Theil and Nagar's analysis, while the price elasticity has a somewhat smaller absolute magnitude.

Since the noise apparently is slightly explosive, the mean $\theta_1$ is not a stable level (having allowed for the independent variables), but may more descriptively be termed an unstable equilibrium. The direction of the explosion is downwards, which may be interpreted as a development in time of consumer habits away from spirits (since logged data are analyzed the explosion is headed towards zero consumption).

Whether it would have been appropriate to difference the series before subjecting them to analysis is an unanswered question at this point. It shall be returned to in Chapter 4, and specifically we shall argue in the Appendix of that chapter, that differencing of this data is not desirable.

## 3.6 Monte Carlo study.

In this section the relative merits of the testing procedures associated with $d$, $\Lambda_{\phi\circ}, t_{\phi\circ}$ and $\tau_{\phi\circ}$ are assessed by means of a simulation study. Although $t_{\phi\circ}$ and $\tau_{\phi\circ}$ are derived at on Bayesian arguments, and therefore do not require any justification by sampling properties, we shall here study how they behave in repeated sampling.

The data generating models, A and B, employed in the present study are

A: $$y_i = \theta_1 + \theta_2 x_{i,2} + e_i \; ; \quad i = 1,2,\ldots,n \qquad (3.67)$$

with $\theta_1 = 10. \; ; \; \theta_2 = 2.$

and where $x_{i,2}$ is a random walk

$$x_{i,2} = \sum_{j=1}^{i} a_j \qquad (3.68)$$

B: $$y_i = \theta x_{i,1} + e_i \; ; \qquad i = 1,2,\ldots,n \qquad (3.69)$$

with $\theta = 2$.

and where $x_{i,1}$ is a series of random deviates

$$x_{i,1} = a_i \qquad (3.70)$$

and $e_i$ is AR-1 noise:

$$\begin{cases} e_1 = M + \varepsilon_1 \\ e_i = e_{i-1} + \varepsilon_i \end{cases} . \qquad (3.71)$$

The $\varepsilon_i$'s of (3.71) and the $a_i$'s of (3.68) and (3.70) are i.i.d. N(0,1). M was set equal to zero for $\phi \geq 1$, while for $\phi < 1$,

$$M = ((1-\phi^2)^{-1/2} - 1)\varepsilon_1 , \qquad (3.72)$$

i.e. the noise follows a reversible stationary scheme ((2.5) and (2.6) in Chapter 2) when $\phi < 1$.

It is easy to visualize many practical circumstances where models like A and B would emerge, at least as a first step toward move comprehensive process descriptions. For instance Model A might describe economic relationships like those examined in Sections 3.5.2 and 3.5.3. As Model A stands (3.67), $\theta_2$ is an elasticity equal to 2., and the "shocks", $a_i$, in the independent variable $x_{i,2}$ have a variance equal to that of $\varepsilon_i$ associated with the noise. But as far as $\phi$ is concerned the data generated by A could equally well have originated from a process with $\theta_2 = 1$, say, and $\mathrm{Var}(a_i) = 4$. In any event the impact of the independent variable on $y_i$ is more pronounced than that of the noise, although the two components are comparable.

A model like B might turn up under similar circumstances if it is appropriate to analyze differenced data, as is commonly done in economics.

From each model 400 independent samples were generated for sample sizes of $n = 15,25,60$ and for each $\phi = -.5,-.4,\ldots,1.1,1.2$, except for



the combination $n = 60$ and $\phi = 1.2$; i.e. a total of $400 \times (2 \times 18 + 1 \times 17)$ independent samples for each model. 400 additional samples were generated by each model with $\phi = 0$ and for $n = 15,25,60$. The models A and B generated data in parallel, i.e. they employed the same $a_i$'s and $\varepsilon_i$'s for corresponding samples (same $\phi$, same $n$). Essentially all calculations were done in double percision.

### 3.6.1 Power curves.

Figures 3.8 (for $n = 15$), 3.9 (for $n = 25$) and 3.10 (for $n = 60$) display the empirical power curves of the alternative testing procedures, d, $\Lambda_0$, $t_0$ and $\tau_0$ relating to the hypothesis $\phi = 0$. The curves show powers corresponding to a theoretical size $\alpha = .05$ for two-sided test; the one-sided power curves are given in the appendix.

The ordinate scale in Figures 3.8, 3.9 and 3.10 is linear in Arcsin$\sqrt{\text{power}}$, going from 0° to 90°. In this scaling $\alpha = .05 \sim 12.9°$, and the standard deviation of all points is approximately 1.4°, except when $\phi = 0$ where it is about 1.°.

If the DW-test is credited with the power of $d_L$, then d still performs comparatively well in relation to Model A, in particular for $.0 < \phi < .9$. For negative and very large positive $\phi$, the alternatives $t_0$ and $\tau_0$ show greater power. In the context of Model B $t_0$ and $\tau_0$ are superior to d. This relative deterioration demonstrates, that the mentioned shortcomings of the DW-test for models lacking a mean are substantive. It is noted that $t_0$ and $\tau_0$ behave remarkably alike.

Throughout $\Lambda_0$ is seemingly more powerful than $t_0$ and $\tau_0$, however this appears attributable to the fact that the size of the $\Lambda_0$-test is actually significantly larger than the sizes of $t_0$ and $\tau_0$, which as a whole do not differ significantly from $\alpha = .05$. In fact the $\Lambda_0$-curve

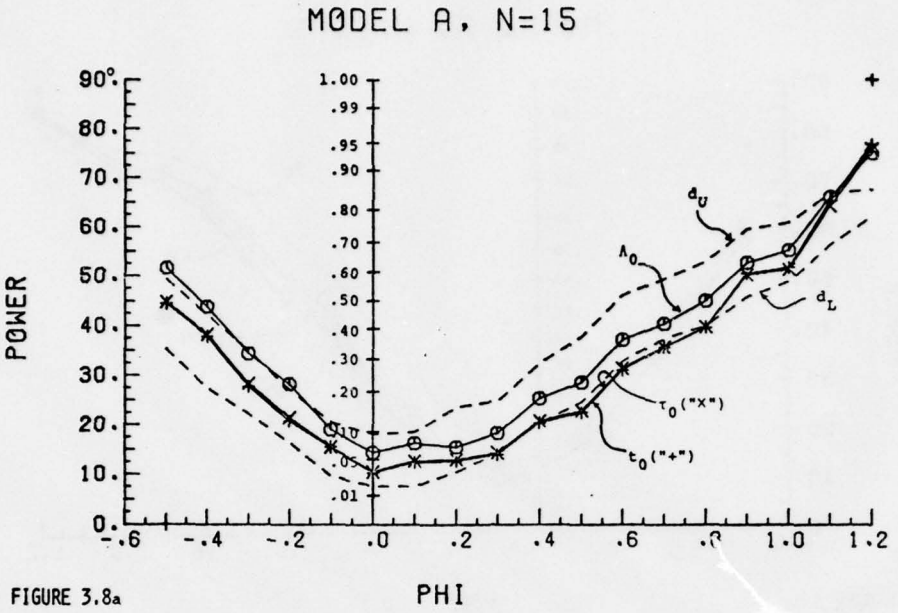



FIGURE 3.8a



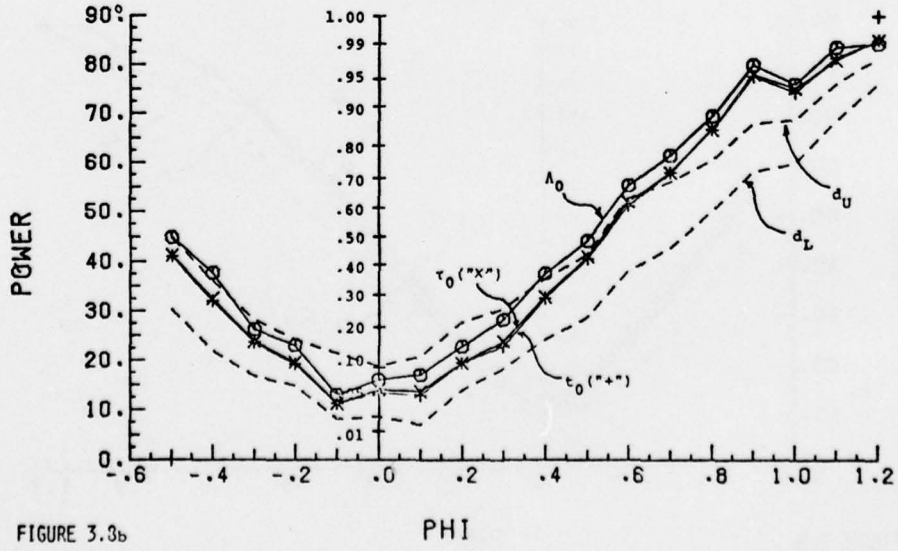



FIGURE 3.8b

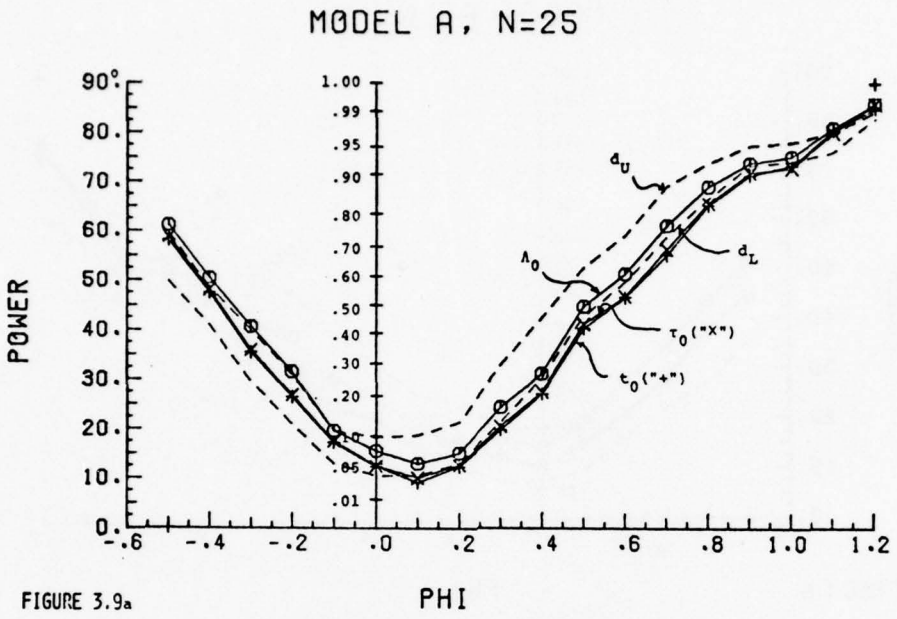



FIGURE 3.9a



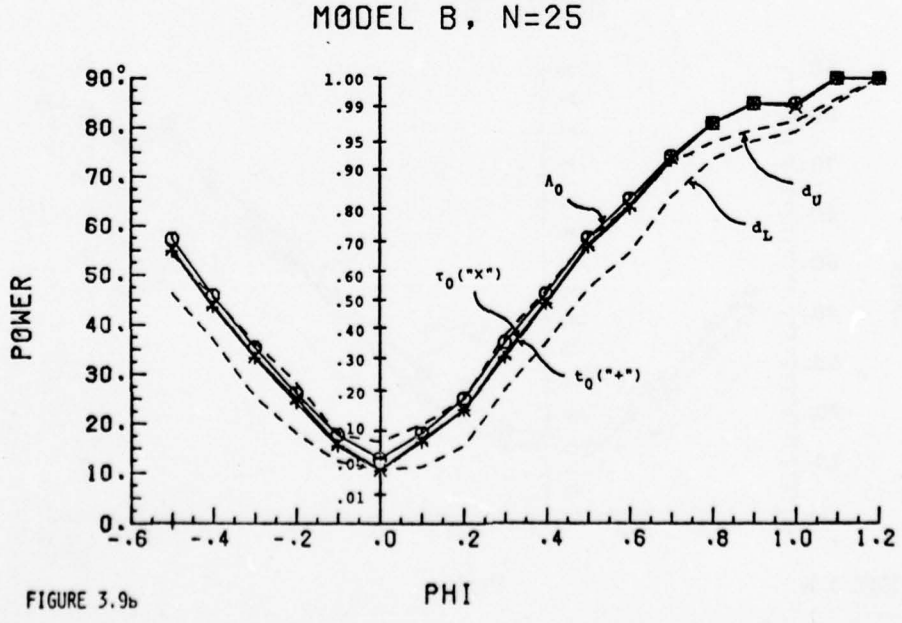



FIGURE 3.9b

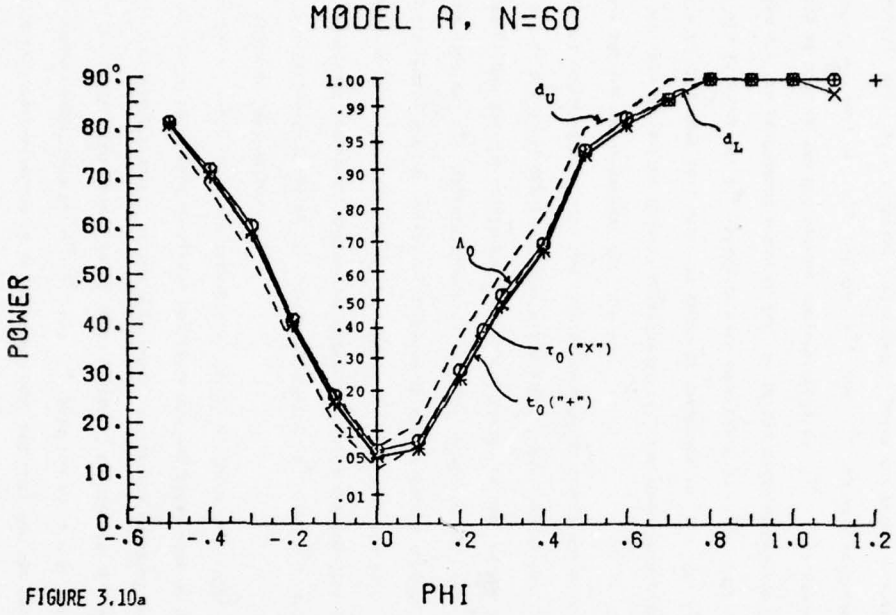



FIGURE 3.10a



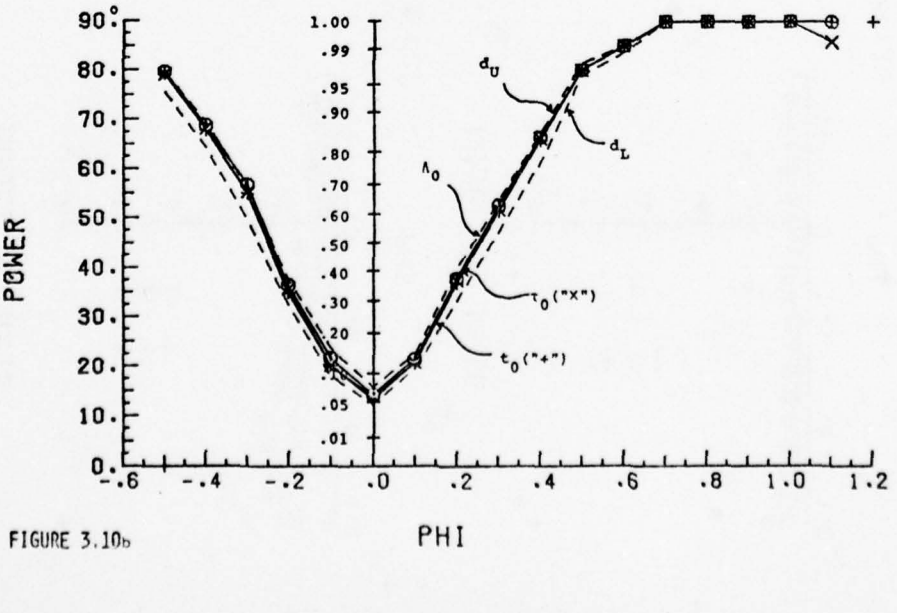



FIGURE 3.10b

looks much like those of $t_0$ and $\tau_0$, just shifted a bit upwards.

3.6.2 Size curves.

The size consideration is an important one, not only for the testing procedures symbolized by $\Lambda_0$, $t_0$ and $\tau_0$ relating to $\phi = 0$, but equally for their counterparts relating to a general hypothesis $\phi = \phi^\circ$. Actually the appropriateness of the approximate confidence interval defined by (3.47) is checked by seeing how close the empirical size of the $\chi^2$-test for $\phi = \phi^\circ$ is to its theoretical value $\alpha$ (here $\alpha = .05$), over a range of $\phi^\circ$ values.

In relation to Model B, the empirical sizes of $t_{\phi^\circ}$ and $\tau_{\phi^\circ}$ are almost identical throughout and do not differ significantly from the target value of $\alpha = .05$. For large $\phi^\circ$, where the approximately non-informative prior for $\phi$ starts to distinguish itself from a uniform one, the size of $\tau_{\phi^\circ}$ appears closer to $\alpha = .05$ than that of $t_{\phi^\circ}$. The size graphs for the likelihood ratio based tests are also rather flat when $\phi^\circ < 1$, but their empirical levels lie significantly above the theoretical value of $\alpha = .05$, the displacement being most pronounced for $n = 15$, but still present for $n = 60$.

All size curves for Model A increase as $\phi^\circ$ increases to a peak at $\phi^\circ = 1$, then they fall off. In terms of closeness to $\alpha = .05$, $\Lambda_{\phi^\circ}$ is clearly inferior to $t_{\phi^\circ}$, which in turn must yield to $\tau_{\phi^\circ}$. This explains why the confidence interval for $\phi$ in the examples Section 3.5 were not as wide as the HPD regions corresponding to $\tau_{\phi^\circ}$. The trouble when $\phi^\circ = 1$ not only for $\Lambda_{\phi^\circ}$ but for $t_{\phi^\circ}$ and $\tau_{\phi^\circ}$ as well confirms empirically, that $\phi^\circ = 1$ being a singularity point for Model A cannot be handled like other points.



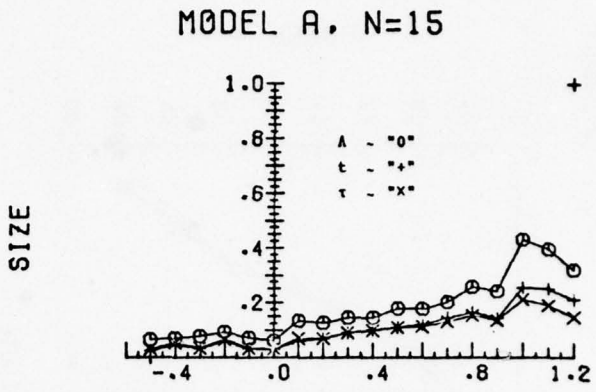



FIGURE 3.11a

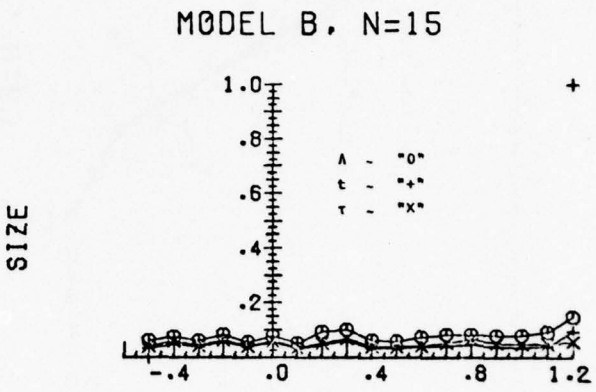



FIGURE 3.11b

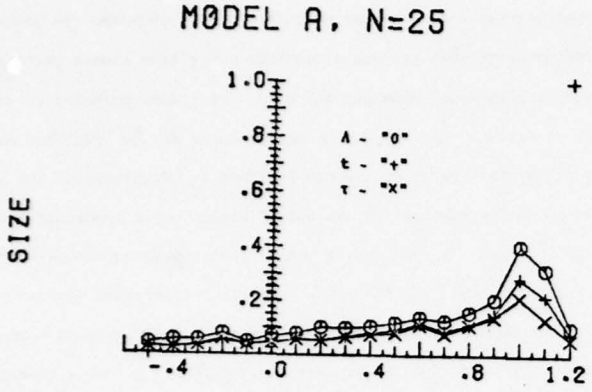



FIGURE 3.12a

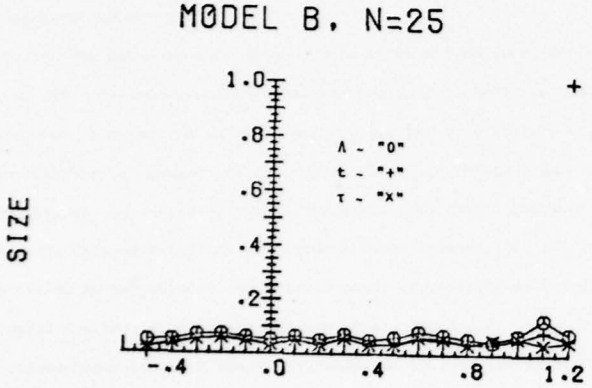



FIGURE 3.12b



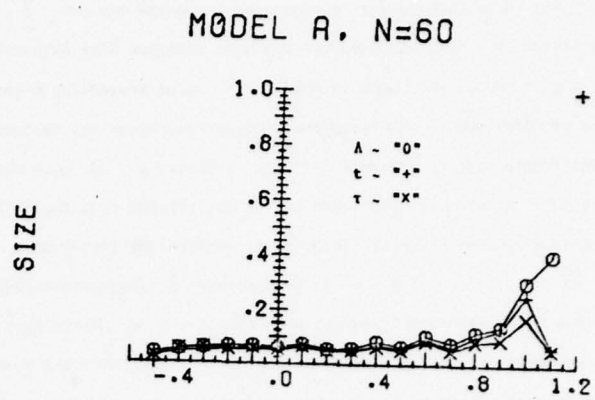



FIGURE 3.13a

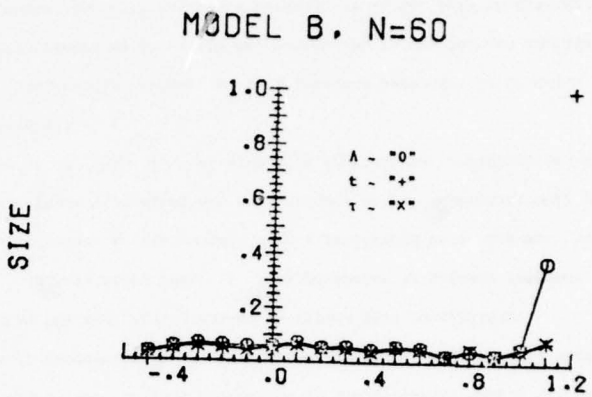



FIGURE 3.13b

## 3.7 Conclusion.

Least squares regression is often applied in situations where it is feared that the observations are serially correlated. Since the usual techniques of analysis rest upon an assumption of independence it is a common practice under these circumstances to test for autocorrelation using the well known Durbin-Watson test. (This test is equivalent to obtaining an estimate $\check{\phi}$ of a first order autoregressive parameter $\phi$ from the residuals left by an ordinary least squares fit, and then seeing whether $\check{\phi}$ is significantly different from zero. Because the distribution of $\check{\phi}$ depends on the matrix $\underset{\sim}{X}$ of independent variables, this is done by referring to bound for the desired significance point.)

In this Chapter some questions surrounding this practice have been approached from two points of view, namely using a likelihood approach and a Bayesian approach.

First the question may be raised whether the DW-test can be improved upon. A direct likelihood ratio approach yields an approximate significance test, $\Lambda_0$, which, being asymptotic only, does not involve $\underset{\sim}{X}$. From a Bayesian point of view one can regard the inclusion or exclusion of some hypothesized value $\phi = \phi^\circ$ (=0) in, say, the $100(1-\alpha)\%$ HPD region as a statistical test. Test criterions $t_0$ and $\tau_0$ of this kind were produced, which do not depend on $\underset{\sim}{X}$; and while they do not require any justification by sampling properties it was studied by simulation how they and $\Lambda_0$ as alternatives to the DW-test, compare in terms of power in repeated sampling. For a studied model form involving a mean, the empirical powers were found comparable for all four alternatives; but the DW-test was somewhat less powerful in relation to a studied model form without a mean.



Secondly when serial dependence is to be expected, it may be questioned whether it actually makes sense to test a null hypothesis of independence. The DW-test seems often to have been abused to justify a convenient but unlikely assumption of independence, mainly for the purpose of proceeding as if, by a stroke of magic, independence were created by the test having failed to reject this possibility.

In situations where $\phi$ is a parameter of primary interest, inference about $\phi$ may be made from a likelihood or a Bayesian point of view. These approaches not only produce better (unconditional) estimates of $\phi$ than $\check{\phi}$, but also give approximate confidence or HPD intervals for $\phi$.

Ordinarily however $\phi$ is a nuisance parameter in relation to $\underset{\sim}{\theta}$. As illustrated by the examples, proceeding in the fashion implied by the DW testing approach carries a penalty, as it may lead to drawing wrong inferences about the primary parameters $\underset{\sim}{\theta}$, both in regard to location and to precision. This may happen when $\phi$ is different from zero but not detected significantly so, and inferences about $\underset{\sim}{\theta}$ subsequently are made conditional on $\phi = 0$. Or even if serial correlation is detected, but inferences about $\underset{\sim}{\theta}$ are made as if $\phi = \check{\phi}$.

When serial correlation is expected, it would seem more sensible to allow for this possibility in the model and estimate $\phi$ simultaneously with $\underset{\sim}{\theta}$. Minimizing $SS(\underset{\sim}{\theta},\phi)$, Equation (3.10), with respect to $\phi$ produces the maximum likelihood estimate, $\hat{\phi}$. From a sampling theory viewpoint inferences about $\underset{\sim}{\theta}$ may now be made conditional on $\phi = \hat{\phi}$ by substituting this estimate into the independence inducing transformation, $\underset{\sim}{c}$, and analyzing the transformed linear model using the usual techniques. From the Bayesian viewpoint this parallels approximation

of the marginal posterior distribution of $\underset{\sim}{\theta}$ by the posterior distribution of $\underset{\sim}{\theta}$ conditional on $\phi = \hat{\phi}$.

Appendix. Additional results from the Monte Carlo study.

Figures 3A.1, 3A.2 and 3A.3 show the empirical onesided power curves of the alternative testing procedures, d, $t_0$ and $\tau_0$ relating to the hypothesis $\phi = 0$, and at the theoretical level $\alpha = .05$.

Tables 3A.I, 3A.II and 3A.III list the averages and standard deviations, s , around the averages, of the estimators $\check{\phi}$ , $\hat{\phi}$ and $\hat{\hat{\phi}}$ computed from the 400 replications for each model, A and B, for each sample size n and each $\phi$-value. Overall $\hat{\phi}$ is clearly a much better estimator of $\phi$ than $\check{\phi}$ . In the Bayesian analysis the estimator $\hat{\hat{\phi}}$ plays no prominent role except as a partial description of the marginal posterior distribution of $\phi$ . The difference between $\hat{\hat{\phi}}$ and $\hat{\phi}$ is caused by the noninformative prior for $\phi$ (relative to a uniform one). It is noted, that on the average it does not make $\overset{*}{\phi}$ overshoot $\phi$ .



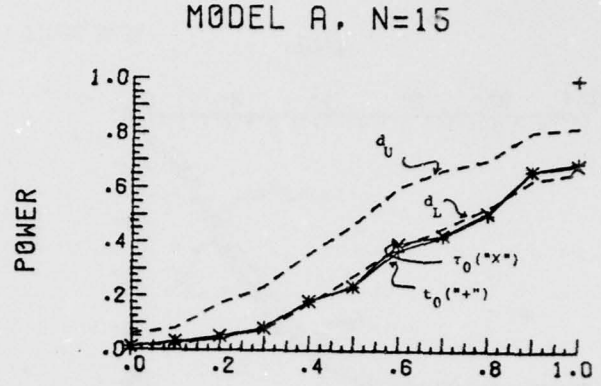



FIGURE 3A.1a

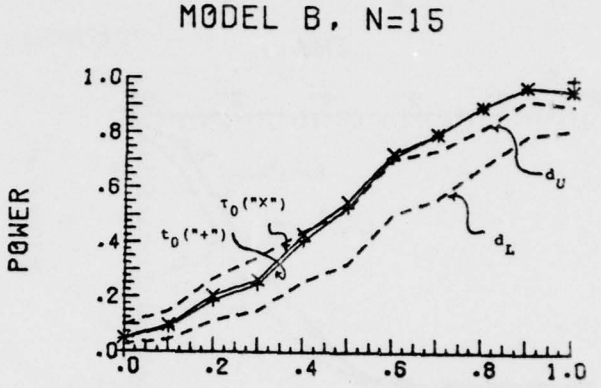



FIGURE 3A.1b

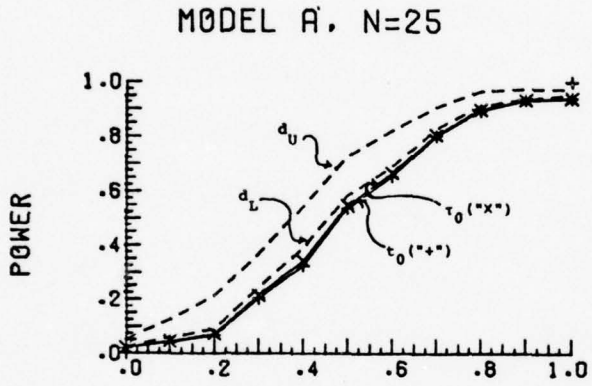



FIGURE 3A.2a

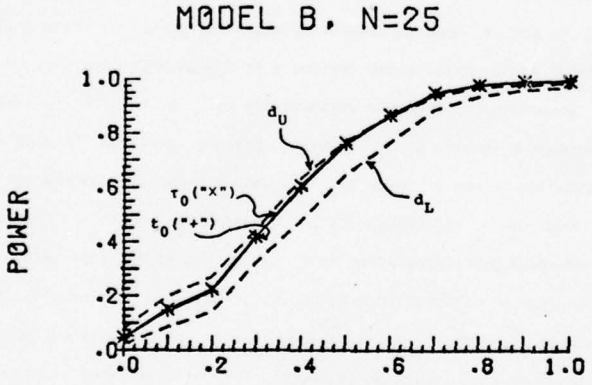



FIGURE 3A.2b



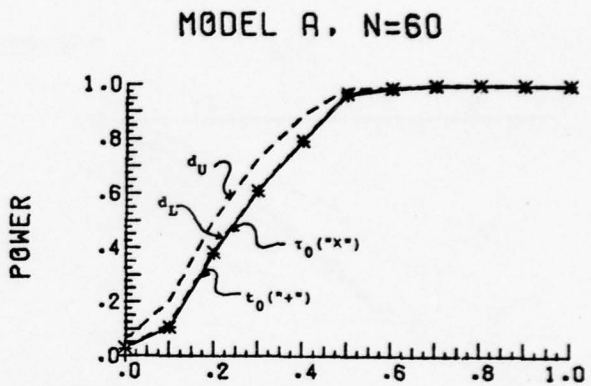



FIGURE 3A.3a

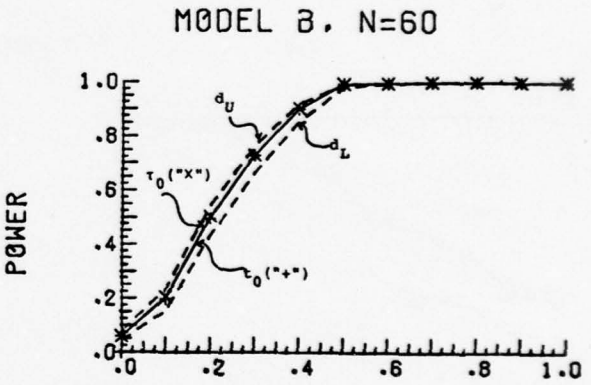



FIGURE 3A.3b

TABLE 3A.I

| Model | $\phi$ | n | $\check{\phi}$ | s | $\hat{\phi}$ | s | $\hat{\hat{\phi}}$ | s |
|---|---|---|---|---|---|---|---|---|
| A | -.50 | 15 | -.410 | .216 | -.514 | .216 | -.562 | .242 |
| | -.40 | 15 | -.342 | .219 | -.451 | .232 | -.496 | .269 |
| | -.30 | 15 | -.270 | .221 | -.382 | .229 | -.398 | .261 |
| | -.20 | 15 | -.185 | .235 | -.271 | .263 | -.297 | .299 |
| | -.10 | 15 | -.110 | .227 | -.187 | .253 | -.205 | .280 |
| | .00 | 15 | -.040 | .225 | -.110 | .254 | -.120 | .277 |
| | .10 | 15 | .001 | .240 | -.054 | .281 | -.057 | .310 |
| | .20 | 15 | .082 | .255 | .028 | .285 | .038 | .337 |
| | .30 | 15 | .141 | .247 | .095 | .290 | .109 | .325 |
| | .40 | 15 | .212 | .253 | .185 | .305 | .207 | .342 |
| | .50 | 15 | .273 | .251 | .252 | .306 | .286 | .356 |
| | .60 | 15 | .362 | .244 | .357 | .301 | .406 | .365 |
| | .70 | 15 | .402 | .258 | .401 | .308 | .443 | .344 |
| | .80 | 15 | .432 | .251 | .449 | .305 | .510 | .424 |
| | .90 | 15 | .512 | .234 | .587 | .293 | .662 | .369 |
| | 1.00 | 15 | .528 | .243 | .614 | .304 | .673 | .344 |
| | 1.10 | 15 | .628 | .249 | .824 | .300 | .895 | .397 |
| | 1.20 | 15 | .671 | .234 | 1.045 | .255 | 1.096 | .272 |
| B | -.50 | 15 | -.360 | .225 | -.457 | .243 | -.496 | .267 |
| | -.40 | 15 | -.252 | .229 | -.350 | .261 | -.415 | .290 |
| | -.30 | 15 | -.205 | .225 | -.261 | .271 | -.284 | .298 |
| | -.20 | 15 | -.131 | .248 | -.170 | .290 | -.185 | .317 |
| | -.10 | 15 | -.055 | .237 | -.064 | .268 | -.070 | .293 |
| | .00 | 15 | .014 | .235 | .006 | .281 | .007 | .308 |
| | .10 | 15 | .066 | .234 | .102 | .273 | .111 | .298 |
| | .20 | 15 | .151 | .254 | .197 | .281 | .215 | .306 |
| | .30 | 15 | .203 | .245 | .267 | .276 | .291 | .303 |
| | .40 | 15 | .270 | .249 | .364 | .267 | .495 | .294 |
| | .50 | 15 | .333 | .229 | .451 | .242 | .492 | .270 |
| | .60 | 15 | .429 | .221 | .553 | .241 | .602 | .269 |
| | .70 | 15 | .481 | .252 | .616 | .245 | .669 | .273 |
| | .80 | 15 | .517 | .227 | .715 | .206 | .771 | .231 |
| | .90 | 15 | .610 | .201 | .837 | .163 | .889 | .176 |
| | 1.00 | 5 | .637 | .211 | .902 | .186 | .969 | .213 |
| | 1.10 | 15 | .749 | .207 | 1.026 | .160 | 1.077 | .168 |
| | 1.20 | 15 | .824 | .156 | 1.136 | .154 | 1.179 | .190 |



TABLE 3A.II

| Model | $\phi$ | n | $\check{\phi}$ | s | $\hat{\phi}$ | s | $\hat{\hat{\phi}}$ | s |
|---|---|---|---|---|---|---|---|---|
| A | -.50 | 25 | -.438 | .173 | -.504 | .166 | -.528 | .175 |
| | -.40 | 25 | -.360 | .178 | -.419 | .181 | -.438 | .190 |
| | -.30 | 25 | -.272 | .183 | -.330 | .183 | -.346 | .192 |
| | -.20 | 25 | -.182 | .204 | -.233 | .206 | -.244 | .216 |
| | -.10 | 25 | -.108 | .190 | -.157 | .194 | -.165 | .203 |
| | .00 | 25 | -.045 | .202 | -.083 | .212 | -.086 | .222 |
| | .10 | 25 | .044 | .196 | .008 | .210 | .009 | .221 |
| | .20 | 25 | .126 | .183 | .098 | .198 | .104 | .209 |
| | .30 | 25 | .201 | .191 | .176 | .209 | .186 | .221 |
| | .40 | 25 | .283 | .183 | .264 | .206 | .278 | .217 |
| | .50 | 25 | .367 | .184 | .359 | .210 | .378 | .222 |
| | .60 | 25 | .434 | .181 | .437 | .204 | .460 | .215 |
| | .70 | 25 | .510 | .171 | .523 | .193 | .554 | .209 |
| | .80 | 25 | .580 | .164 | .612 | .193 | .655 | .225 |
| | .90 | 25 | .654 | .171 | .712 | .197 | .768 | .253 |
| | 1.00 | 25 | .689 | .174 | .753 | .200 | .800 | .225 |
| | 1.10 | 25 | .784 | .182 | .987 | .167 | 1.030 | .211 |
| | 1.20 | 25 | .840 | .145 | 1.177 | .097 | 1.187 | .094 |
| B | -.50 | 25 | -.403 | .175 | -.474 | .181 | -.496 | .190 |
| | -.40 | 25 | -.333 | .178 | -.385 | .189 | -.403 | .199 |
| | -.30 | 25 | -.242 | .190 | -.283 | .203 | -.296 | .212 |
| | -.20 | 25 | -.147 | .208 | -.178 | .223 | -.187 | .233 |
| | -.10 | 25 | -.076 | .192 | -.095 | .209 | -.100 | .220 |
| | .00 | 25 | -.010 | .200 | -.010 | .216 | -.011 | .227 |
| | .10 | 25 | .080 | .199 | .094 | .226 | .098 | .236 |
| | .20 | 25 | .159 | .183 | .186 | .203 | .195 | .213 |
| | .30 | 25 | .247 | .191 | .292 | .201 | .305 | .210 |
| | .40 | 25 | .323 | .182 | .381 | .185 | .399 | .194 |
| | .50 | 25 | .405 | .179 | .474 | .182 | .496 | .191 |
| | .60 | 25 | .475 | .178 | .556 | .183 | .582 | .192 |
| | .70 | 25 | .564 | .166 | .661 | .164 | .692 | .173 |
| | .80 | 25 | .638 | .155 | .744 | .143 | .780 | .155 |
| | .90 | 25 | .718 | .157 | .856 | .122 | .896 | .146 |
| | 1.00 | 25 | .750 | .152 | .925 | .123 | .968 | .135 |
| | 1.10 | 25 | .887 | .106 | 1.057 | .093 | 1.085 | .094 |
| | 1.20 | 25 | .929 | .068 | 1.188 | .053 | 1.196 | .054 |

TABLE 3A.III

| Model | $\phi$ | n | $\check{\phi}$ | s | $\hat{\phi}$ | s | $\hat{\hat{\phi}}$ | s |
|---|---|---|---|---|---|---|---|---|
| A | -.50 | 60 | -.476 | .106 | -.500 | .107 | -.510 | .109 |
| | -.40 | 60 | -.382 | .121 | -.408 | .120 | -.416 | .122 |
| | -.30 | 60 | -.305 | .122 | -.328 | .122 | -.334 | .125 |
| | -.20 | 60 | -.206 | .127 | -.228 | .128 | -.232 | .131 |
| | -.10 | 60 | -.104 | .132 | -.122 | .134 | -.124 | .137 |
| | .00 | 60 | -.018 | .121 | -.035 | .123 | -.035 | .125 |
| | .10 | 60 | .075 | .132 | .061 | .133 | .062 | .136 |
| | .20 | 60 | .186 | .119 | .174 | .123 | .177 | .125 |
| | .30 | 60 | .259 | .121 | .250 | .124 | .255 | .127 |
| | .40 | 60 | .336 | .124 | .327 | .127 | .334 | .130 |
| | .50 | 60 | .445 | .117 | .442 | .123 | .451 | .126 |
| | .60 | 60 | .533 | .114 | .535 | .121 | .546 | .125 |
| | .70 | 60 | .624 | .101 | .631 | .107 | .646 | .111 |
| | .80 | 60 | .718 | .099 | .729 | .105 | .748 | .110 |
| | .90 | 60 | .790 | .069 | .818 | .092 | .841 | .097 |
| | 1.00 | 60 | .857 | .090 | .907 | .088 | .939 | .115 |
| | 1.10 | 60 | .942 | .059 | 1.095 | .029 | 1.098 | .026 |
| B | -.50 | 60 | -.462 | .110 | -.489 | .111 | -.498 | .113 |
| | -.40 | 60 | -.366 | .121 | -.391 | .124 | -.399 | .127 |
| | -.30 | 60 | -.289 | .122 | -.310 | .128 | -.316 | .131 |
| | -.20 | 60 | -.191 | .126 | -.202 | .134 | -.205 | .136 |
| | -.10 | 60 | -.089 | .132 | -.092 | .134 | -.094 | .137 |
| | .00 | 60 | -.001 | .121 | .001 | .127 | .001 | .129 |
| | .10 | 60 | .091 | .131 | .099 | .138 | .101 | .141 |
| | .20 | 60 | .199 | .119 | .214 | .124 | .218 | .126 |
| | .30 | 60 | .275 | .123 | .291 | .125 | .297 | .127 |
| | .40 | 60 | .354 | .119 | .378 | .122 | .385 | .124 |
| | .50 | 60 | .464 | .115 | .493 | .117 | .502 | .119 |
| | .60 | 60 | .553 | .109 | .585 | .109 | .597 | .112 |
| | .70 | 60 | .644 | .096 | .683 | .092 | .699 | .095 |
| | .80 | 60 | .737 | .091 | .781 | .085 | .800 | .089 |
| | .90 | 60 | .819 | .083 | .875 | .068 | .896 | .067 |
| | 1.00 | 60 | .901 | .064 | .972 | .049 | .995 | .057 |
| | 1.10 | 60 | .972 | .032 | 1.097 | .022 | 1.099 | .021 |



## CHAPTER 4

## To Difference or not to Difference Data Series in Linear Model Analysis: A Bayesian Study.

For the purpose of definiteness, consider some data analyzed by Coen, Gomme & Kendall [1969] (CGK for short). In an attempt to forecast the Financial Times ordinary share index, they relate this dependent variable, $y_i$, detrended, to two detrended "leading" independent variables $x_{1,i}$ and $x_{2,i}$, viz. U.K. car production and the Financial Times commodity index (their equation number 7):

$$y_i = \theta_1 + \theta_2 x_{1,i} + \theta_3 x_{2,i} + \theta_4 i + N_i \qquad (4.1)$$

where $x_{1,i}$ is the observed value 6 time periods (quarters) previous to $y_i$, and $x_{2,i}$ has a lead time of 7 periods.

This data set raises a number of questions. First of all whether the noise $N_i$ might be serially correlated and not independent as the CGK analysis tacitly assumed. Secondly how this serial correlation, if indeed present, affects the inferences about the regression parameters $\theta_j$, where CGK's estimates of $\theta_2$ and $\theta_3$ were found highly significantly different from zero.

Both these questions were addressed by Box and Newbold [1971] (B&N), who demonstrated, that the noise was in fact highly serially dependent, and that taking the autocorrelation into account the claimed dependencies of $y$ on $x_1$ and $x_2$ disappeared. B&N entertained several alternative noise structures, the most prominent of which were a first order autoregressive (AR-1) model, and a first order integrated moving average (IMA-1) model (i.e. a noise model for which the first difference is a first order moving average process). This latter model fits the data only slightly less well than the AR-1 model, but B&N did not

commit themselves to any preferred form over the other, since it made little difference in demonstrating their point. This does however leave a third question, namely whether differencing should be carried out or not.

Writing the linear model

$$\underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times p}{\underset{\approx}{X}}\ \underset{p\times 1}{\underset{\sim}{\theta}} + \underset{n\times 1}{\underset{\sim}{N}} \qquad (4.2)$$

the noise $\underset{\sim}{N}$ may in great generality be considered generated by an autoregressive integrated moving average (ARIMA) process of order $(p,d,q)$, see Box and Jenkins [1970]. The presence of the deterministic linear model parameters $\underset{\sim}{\theta}$ does not impede a time series modelling of the noise, as the linear and the time series parameters may be estimated simultaneously, and the latter are (asymptotically) distributed as if the deterministic model part did not exist. In such a general approach, the question of differencing takes the form of deciding the order $d$ in the ARIMA model.

The scope of the present investigation is limited to situations, where it is impossible (due to a relatively small number of observations), impractical or perhaps unnecessary to employ more than one first order autoregressive parameter to allow for autocorrelation.

How such a model may be analyzed was the topic of Chapter 2, specifically it was studied how Bayesian inference is made about the model parameters $\underset{\sim}{\theta}$ and $\phi_0$ when the noise $e_i$ of the observations (original or perhaps differenced) follows the AR-1 scheme

$$\begin{cases} e_1 = M_0 + \varepsilon_1 \\ e_i = \phi_0 e_{i-1} + \varepsilon_i \qquad i = 2,3,\ldots,n \end{cases} \qquad (4.3)$$



where $\phi_0$ is the autogressive parameter, $M_0$ is a starting parameter included to accommodate explosive, $(1 < \phi_0)$ as well as stationary situations and the $\varepsilon_i$'s are i.i.d. $N(0,\sigma^2)$, (also independent of the p series of predictor variables $\underset{\approx}{X}$ ).

Two ways of assessing whether or not differencing appears justified in light of the data will now be explored from a Bayesian point of view.

Within the AR-1 noise structure, differencing corresponds to a particular choice of the autoregressive parameter $\phi_0$, namely $\phi_0 = 1$, so it may be wondered whether the data support this value (Section 4.2).

Since the question of differencing is intimately related to the existence (or lack) of a fixed mean for the noise, it may usefully be regarded as one of determining whether one degree of freedom is better *expended on differencing the series (effectively loosing one observation)* rather than on including a mean, while in either case serial correlation is being allowed for by one parameter. In other words, the testimony of the data concerning the appropriateness of differencing here takes the form of relative support of two alternative models with a common deterministic part but with different noise models. (Section 4.3).

The CGK data are reanalyzed in detail (Section 4.4) in particular *the results concerning whether to difference or not are illustrated.*

## 4.1 Joint posterior distribution of $\{\theta, \phi_0\}$.

Analyzing data sets like the CGK data from a Bayesian point of view makes it very clear, why misleading conclusions may be reached about the linear parameters $\underset{\sim}{\theta}$ when serial correlation is ignored.

Assuming that the data $\underset{\sim}{y}$ may adequately be represented by the linear model

$$\underset{\sim}{y} = \underset{\sim}{1}\,\theta_1 + \underset{\approx}{X}_{(1)}\underset{\sim}{\theta}_{(1)} + \underset{\sim}{e} \qquad (4.4)$$

where

$$\underset{n\times p}{\underset{\approx}{X}} = [\underset{\sim}{1}\ \underset{\approx}{X}_{(1)}]\ ;\quad \underset{p\times 1}{\underset{\sim}{\theta}} = \begin{bmatrix}\theta_1\\ \underset{\sim}{\theta}_{(1)}\end{bmatrix} \qquad (4.5)$$

with $\operatorname{rank}(\underset{\approx}{X}) = p$, and where $\underset{\sim}{e}$ follows the AR-1 scheme (4.3); then it was seen in Chapter 2, that the density function for $\underset{\sim}{y}$ is:

$$p(\underset{\sim}{y}|M_0,\underset{\sim}{\theta},\phi_0,\sigma) = (2\pi\sigma^2)^{-n/2}\exp(-\tfrac{1}{2}\sigma^{-2}((\underset{\sim}{y}-\underset{\approx}{X}\underset{\sim}{\theta})'\underset{\approx 0}{C}(\underset{\sim}{y}-\underset{\approx}{X}\underset{\sim}{\theta})-(y_1-\underset{\sim}{x}_1'\underset{\sim}{\theta}-M_0)^2))$$

$$= (2\pi\sigma^2)^{-n/2}\exp(-\tfrac{1}{2}\sigma^{-2}(SS_0(\underset{\sim}{\theta},\phi_0)+(e_1-M_0)^2)) \qquad (4.6)$$

where

$$\underset{\approx 0}{C} = \underset{\approx 0}{c}'\underset{\approx 0}{c}\ ;\quad \underset{(n-1)\times n}{\underset{\approx 0}{c}} = \begin{bmatrix} -\phi_0 & 1 & & 0\\ & \ddots & \ddots & \\ 0 & & -\phi_0 & 1\end{bmatrix} \qquad (4.7)$$

so that

$$SS_0(\underset{\sim}{\theta},\phi) = (\underset{\approx 0}{c}\underset{\sim}{y} - \underset{\approx 0}{c}\underset{\approx}{X}\underset{\sim}{\theta})'(\underset{\approx 0}{c}\underset{\sim}{y} - \underset{\approx 0}{c}\underset{\approx}{X}\underset{\sim}{\theta}) \qquad (4.8)$$

(and where $\underset{\sim}{x}_1'$ is the first row of $\underset{\approx}{X}$). The subscript "$_0$" is used to indicate, that the AR-1 noise relates to the "not differenced" data. Of course the expression (4.6) is the likelihood function for the parameters $\underset{\sim}{\psi} = \{M_0,\underset{\sim}{\theta},\phi_0,\sigma\}$ of the model (4.4) when $\underset{\sim}{y}$ is considered known. It is noted, that maximizing the likelihood function is equivalent to minimizing the (residual) sum of squares of the estimated $e_i$'s, $SS_0(\underset{\sim}{\theta},\phi_0)$; and the minimum is $SS_0(\hat{\underset{\sim}{\theta}},\hat{\phi}_0)$.

It was argued in Chapter 2, that an approximately noninformative prior distribution for $\underset{\sim}{\psi}$ is of the form:

$$p(M_0,\underset{\sim}{\theta},\phi_0,\sigma) \propto |\underset{\approx}{X}'\underset{\approx 0}{C}\underset{\approx}{X}|^{1/2}\,\sigma^{-1}\,p(\phi_0) \qquad (4.9)$$

where $p(\phi_0)$ is a locally noninformative prior for $\phi_0$. The argument leading to (4.9) applies only when $\underset{\approx}{X}'\underset{\approx 0}{C}\underset{\approx}{X}$, or equivalently $\underset{\approx 0}{c}\underset{\approx}{X}$, has



full column rank $p$. Excluding the distinct singularity points $\phi = \phi^*$, if any, where $\operatorname{rank}(\underset{\approx 0}{c}\underset{\approx}{X}) < p$ (in fact equal to $p-1$), the joint posterior for $\{M_0,\underset{\sim}{\theta},\phi_0,\sigma\}$, found by multiplying (4.6) by (4.9), was integrated over $\{M_0,\underset{\sim}{\theta},\sigma\}$ to yield the marginal posterior distribution of $\phi_0$:

$$p(\phi_0|\underset{\sim}{y}) \propto p(\phi_0)(SS_0(\hat{\underset{\sim}{\theta}},\phi_0))^{-\frac{n-p-1}{2}} \qquad (4.10)$$

where

$$\hat{\underset{\sim}{\theta}} = \hat{\underset{\sim}{\theta}}(\phi_0) = (\underset{\approx}{X}'\underset{\approx 0}{C}\underset{\approx}{X})^{-1}\underset{\approx}{X}'\underset{\approx 0}{C}\underset{\sim}{y} \qquad (4.11)$$

Omitting the distinct point(s) $\phi_0 = \phi_0^*$ (if any), (4.10) is normalized as if it was a continuous distribution. Specifically it was suggested in Chapter 2 how $p(\phi_0|\underset{\sim}{y})$ may be approximated by a t-distribution. The exclusion of possible singularity points is only a matter of formality as far as marginal inference about $\underset{\sim}{\theta}$ is concerned.

Conditional on $\phi_0$ ($\neq \phi_0^*$) the distribution of $\underset{\sim}{\theta}$ a posteriori is

$$p_0(\underset{\sim}{\theta}|\phi_0,\underset{\sim}{y}) \sim t_p(\hat{\underset{\sim}{\theta}},s^2(\underset{\approx}{X}'\underset{\approx 0}{C}\underset{\approx}{X})^{-1},\nu) \qquad (4.12)$$

with

$$\nu s^2 = SS_0(\hat{\underset{\sim}{\theta}},\phi_0)\ ;\quad \nu = n-p-1 \qquad (4.13)$$

in particular

$$p_0(\theta_j|\phi_0,\underset{\sim}{y}) \sim t(\hat{\theta}_j,s^2D_{jj},\nu) \qquad (4.14)$$

where $D_{jj}$ comes from

$$[\underset{\approx}{X}\underset{\approx 0}{C}\underset{\approx}{X}]^{-1} = \begin{bmatrix} D_{11} & & & \text{sym}\\ & \ddots & & \\ & & D_{jj} & \\ & & & \ddots\\ D_{1p} & \cdots & & D_{pp}\end{bmatrix} \qquad (4.15)$$

Now the joint posterior distribution of $\{\theta_j,\phi_0\}$ may conveniently be computed as

$$p_0(\theta_j,\phi_0|\underset{\sim}{y}) = p_0(\theta_j|\phi_0,\underset{\sim}{y})\,p(\phi_0|\underset{\sim}{y}) \qquad (4.16)$$

and plotting contours of this joint distribution, $j = 1,2,\ldots,p$, allows a clear appreciation of the dependence of $\theta_j$ on $\phi_0$. The levels of the contours corresponding to specified interior probability contents may be determined numerically. As an approximation it may be used (see for example Box and Tiao [1973] p. 94) that for large n $p(\theta_j,\phi_0|\underset{\sim}{y})$ tends towards normality, so that

$$-2\ell n \frac{p_0(\theta_j,\phi_0)}{p_0(\hat{\theta}_j,\hat{\phi}_0)} \overset{\cdot}{\sim} \chi^2_2 \tag{4.17}$$

which (in the 2 dimensional case) may be expressed as:

$$\frac{p_0(\theta_j^\alpha,\phi_0^\alpha|\underset{\sim}{y})}{p_0(\hat{\theta}_j,\hat{\phi}_0|\underset{\sim}{y})} = e^{-\frac{1}{2}\chi^2_2(\alpha)} = \alpha \tag{4.18}$$

where $\hat{\phi}_0$ is determined by the identity

$$SS_0(\hat{\underset{\sim}{\theta}},\hat{\phi}_0) = \min_{\phi_0} SS(\hat{\underset{\sim}{\theta}},\phi_0) \tag{4.19}$$

and where $(\theta_j^\alpha,\phi_0^\alpha)$ is a point on the boundary of the $100(1-\alpha)\%$ highest posterior density (HPD) region .

4.2 Posterior density of the AR-1 parameter at unity.

For models like (4.4) involving a mean, the point $\phi_0 = 1$ is a singularity point. Thus the plausibility of $\phi_0$ being unity (i.e. whether or not a differencing appears supported by the data) cannot be assessed by looking at $p(\phi_0|\underset{\sim}{y})$ (4.10) as a continuous distribution, since the prior (4.9) does not find justification at this point, and the integration leading to (4.10) does not go through.

The cause of the singularity at $\phi_0 = 1$ is the vanishing of the mean $\theta$, at that point. Specifically (4.4) degenerates to

$$\underset{\sim}{y} = \underset{\approx}{X}_{(1)}\,\underset{\sim}{\theta}_{(1)} + \underset{\sim}{e}(\phi_0 = 1) \tag{4.20}$$

for $\phi_0 = 1$; as namely $e_i(\phi_0 = 1)$, and therefore $y_i$, does not possess a mean. Hence conditional on $\phi_0 = 1$ we have



$$\hat{\underset{\sim}{\theta}}_{(1)} = \left(\underset{\approx}{X}_{(1)}\underset{\approx}{D}\,\underset{\approx}{X}_{(1)}\right)^{-1}\underset{\approx}{X}'_{(1)}\,\underset{\approx}{D}\,\underset{\sim}{y} \tag{4.21}$$

where

$$\underset{\approx}{D} = \underset{\approx}{d}'\underset{\approx}{d}\ ;\ \underset{(n-1)\times n}{\underset{\approx}{d}} = \begin{bmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \tag{4.22}$$

and

$$p_0(\underset{\sim}{\theta}_{(1)}|\phi_0 = 1,\underset{\sim}{y}) \sim t_{p-1}(\hat{\underset{\sim}{\theta}}_{(1)}, s^2(\underset{\approx}{X}'_{(1)}\,\underset{\approx}{D}\,\underset{\approx}{X}_{(1)})^{-1}, \nu) \tag{4.23}$$

with

$$\nu s^2 = SS_0(\hat{\underset{\sim}{\theta}}_{(1)},\phi_0=1) = (\underset{\sim}{dy}-\underset{\approx}{d}\,\underset{\approx}{X}_{(1)}\,\underset{\sim}{y})'(\underset{\sim}{dy}-\underset{\approx}{dX}_{(1)}\underset{\sim}{y});\ \nu = n-p \tag{4.24}$$

It is noted that $SS_0(\hat{\underset{\sim}{\theta}}_{(1)},\phi_0=1)$ is in general not equal to $\lim_{\phi_0 \to 1} SS_0(\hat{\underset{\sim}{\theta}},\phi_0)$. It was seen in Chapter 3 that the latter quantity, which we shall refer to as $SS_0(\hat{\underset{\sim}{\theta}},\phi_0=1)$, may be computed as the residual sum of squares resulting from replacing, when $\phi_0=1$, the redundant parameter $\theta_1$ by a new parameter $\theta_1^*$ serving as a mean for the differenced series, i.e. implying that the model incorporates a deterministic linear trend. So when this artificial trend is not incorporated, then the residual sum of squares, $SS_0$, corresponding to the model (4.4) makes a discontinuous upwards jump at $\phi_0=1$ as a result of the model's loss of one parameter at that point. Thus if the prior for $\phi_0$ is taken to be continuous through $\phi_0=1$ then the posterior density function for $\phi_0$ will not in general be continuous at that point.

In order to determine a prior distribution at $\phi_0 = 1$ for $\{M_0,\underset{\sim}{\theta}_{(1)},\phi_0,\sigma\}$ which is known up to the same multiplicative constant as $p(M_0,\underset{\sim}{\theta},\phi_0,\sigma)$ of (4.9), we shall argue the following way: Conditional on $\sigma$ and $\phi_0$ the prior for the linear parameters $\underset{\sim}{\theta}_{(1)}$ and M is taken locally uniform proportional to k . To determine k we suppose that $\sigma$ is known, and look at the posterior densities of $\phi_0$ at $\phi_0 = 1$ and $\phi_0 = \phi_0'(\neq 1)$. We shall then consider the prior for

$\{M_0, \underset{\sim}{\theta}_{(1)} | \phi_0 = 1\}$ being consistent with $p(M_0, \underset{\sim}{\theta} | \phi_0 = \phi_0')$ in the sense

$$\lim_{\phi_0 \to 1} \; \underset{\phi_0 = \phi_0'}{E} \ln p(\phi_0 = \phi_0' | \underset{\sim}{y}) = \underset{\phi_0 = 1}{E} \ln p(\phi_0 = 1 | \underset{\sim}{y}). \tag{4.25}$$

We find

$$p(\phi_0 = \phi_0' | \underset{\sim}{y}) \propto \int_{M_0, \underset{\sim}{\theta}} p(M_0, \underset{\sim}{\theta} | \phi_0 = \phi_0') p(\phi_0 = \phi_0') p(\underset{\sim}{y} | M_0, \underset{\sim}{\theta}, \phi_0) dM_0 d\underset{\sim}{\theta}$$

$$\propto p(\phi_0 = \phi_0') (2\pi\sigma^2)^{-\frac{n-p-1}{2}} \exp\left(-\tfrac{1}{2}\sigma^{-2} SS_0(\hat{\underset{\sim}{\theta}}, \phi_0 = \phi_0')\right) \tag{4.26}$$

and

$$p(\phi_0 = 1 | \underset{\sim}{y}) \propto \int_{M_0, \underset{\sim}{\theta}_{(1)}} p(M_0, \underset{\sim}{\theta}_{(1)} | \phi_0 = 1) p(\phi_0 = 1) p(\underset{\sim}{y} | M_0, \underset{\sim}{\theta}_{(1)} \phi_0) dM_0 d\underset{\sim}{\theta}_{(1)}$$

$$\propto p(\phi_0 = 1) \left| \underset{\approx}{X}'_{(1)} \underset{\approx}{D} \underset{\approx}{X}_{(1)} \right|^{-1/2} k (2\pi\sigma^2)^{-\frac{n-p}{2}} \exp\left(-\tfrac{1}{2}\sigma^{-2} SS_0(\hat{\underset{\sim}{\theta}}_{(1)}, \phi_0 = 1)\right) \tag{4.27}$$

where it is understood in (4.26) and (4.27), that the proportionality constants are the same. Taking expectation of the logarithm of these expressions and equating gives

$$\ln p(\phi_0 = \phi_0') - \frac{n-p-1}{2} \ln(2\pi\sigma^2) - \tfrac{1}{2}(n-p-1) \tag{4.28}$$

$$= \ln p(\phi_0 = 1) \ln\left( \left| \underset{\approx}{X}'_{(1)} \underset{\approx}{D} \underset{\approx}{X}_{(1)} \right|^{-1/2} k \right) - \frac{n-p}{2} \ln(2\pi\sigma^2) - \tfrac{1}{2}(n-p)$$

Letting $\phi_0' \to 1$ the prior densities at $\phi_0 = \phi_0'$ and $\phi_0 = 1$ cancel, leaving

$$k = \left| \underset{\approx}{X}'_{(1)} \underset{\approx}{D} \underset{\approx}{X}_{(1)} \right|^{1/2} e^{1/2} (2\pi\sigma^2)^{1/2} \tag{4.29}$$

hence we are led to choose

$$p(M_0, \underset{\sim}{\theta}_{(1)}, \phi_0, \sigma) = p(M_0, \underset{\sim}{\theta}_{(1)} | \sigma, \phi_0 = 1)\; p(\sigma)\; p(\phi_0)\big|_{\phi_0 = 1}$$

$$= \left| \underset{\approx}{X}'_{(1)} \underset{\approx}{D} \underset{\approx}{X}_{(1)} \right|^{1/2} e^{1/2} (2\pi\sigma^2)^{1/2} \sigma^{-1} p(\phi_0 = 1). \tag{4.30}$$



The posterior density $p(\phi_0 | \underset{\sim}{y})$ at $\phi_0 = 1$ is now determined from considering

$$\frac{p(\phi_0 = 1 | \underset{\sim}{y})}{p(\phi_0 = \phi_0' | \underset{\sim}{y})} = \frac{p(\phi_0 = 1)}{p(\phi_0 = \phi_0')} \; \frac{\int_{M_0, \underset{\sim}{\theta}_{(1)}, \sigma} I_{\phi_0 = 1}(M_0, \underset{\sim}{\theta}_{(1)}, \sigma) dM_0 d\underset{\sim}{\theta}_{(1)} d\sigma}{\int_{M_0, \underset{\sim}{\theta}, \sigma} I_{\phi_0 = \phi_0'}(M_0, \underset{\sim}{\theta}, \sigma) dM_0 d\underset{\sim}{\theta}\, d\sigma}$$

$$= \frac{p(\phi_0 = 1)}{p(\phi_0 = \phi_0')} \; \frac{e^{1/2} (SS_0(\hat{\underset{\sim}{\theta}}_{(1)}, \phi_0 = 1))^{-\frac{n-p-1}{2}}}{(SS_0(\hat{\underset{\sim}{\theta}}, \phi_0 = \phi_0'))^{-\frac{n-p-1}{2}}} \tag{4.31}$$

with

$$I_{\phi_0 = 1}(M_0, \underset{\sim}{\theta}_{(1)}, \sigma) =$$

$$\left| \underset{\approx}{X}'_{(1)} \underset{\approx}{D} \underset{\approx}{X}_{(1)} \right|^{1/2} e^{1/2} (2\pi\sigma^2)^{-\frac{n-1}{2}} \sigma^{-1} \exp\left(-\tfrac{1}{2}\sigma^{-2} (SS_0(\underset{\sim}{\theta}_{(1)}, \phi_0 = 1) + (e_1 - M_0)^2)\right)$$

and

$$I_{\phi_0 = \phi_0'}(M_0, \underset{\sim}{\theta}, \sigma) =$$

$$\left| \underset{\approx}{X}' \underset{\approx}{C}_0 \underset{\approx}{X} \right|^{1/2} (2\pi\sigma^2)^{-\frac{n}{2}} \sigma^{-1} \exp\left(-\tfrac{1}{2}\sigma^{-2} (SS_0(\hat{\underset{\sim}{\theta}}, \phi_0 = \phi_0') + (e_1 - M_0)^2)\right).$$

Now letting $\phi_0' \to 1$, we find the following expression for the density at $\phi_0 = 1$:

$$p(\phi_0 = 1 | \underset{\sim}{y}) = \lim_{\phi_0 \to 1} p(\phi_0 | \underset{\sim}{y})\; e^{1/2} \left( \frac{SS_0(\hat{\underset{\sim}{\theta}}_{(1)}, \phi_0 = 1)}{SS_0(\hat{\underset{\sim}{\theta}}, \phi_0 = 1)} \right)^{-\frac{n-p-1}{2}} \tag{4.32}$$

The discrete jump of $p(\phi_0 | \underset{\sim}{y})$ at $\phi_0 = 1$ may be upwards or downwards depending on the last factor in (4.32); since it is less than or equal to 1, the largest possible upwards jump is by a factor $e^{1/2} \doteq 1.65$.

Other singularity point $\phi_0 = \phi_0^*$ (if any), may obviously be taken care of similarly.

## 4.3 Expending one degree of freedom on differencing rather than on a fixed mean?

In this section the question of whether some autocorrelated series should perhaps be differenced prior to a linear model analysis is approached from the point of view of determining to what extent the data are more consistent with a model having a fixed mean and an AR-1 noise (4.3), or with a model which includes no mean and has a nonstationary noise component. In the Bayesian framework the testimony of the data concerning this question may be extracted in terms of posterior probabilities for two alternative models with a common deterministic part, but with alternative noise models, both allowing for serial correlation by one parameter.

In this formulation the problem is similar to the one known as *model discrimination*, *Box & Hill [1967]*, *Box & Henson [1969]*, *Kanemasu [1973]*. Their problem was discrimination among alternative deterministic models with identical noise components; in the present problem the discrimination is between alternative noise models for the same deterministic model.

Singularity points require no special attention in this analysis, in particular the model (4.4) involves, with probability one, a fixed mean $\theta_1$.

### 4.3.1 The alternative likelihood function

A natural alternative, $\underset{\sim}{f}$, to the AR-1 noise structure $\underset{\sim}{e}$ (4.3), consists in letting the noise of the differenced series follow an AR-1 process, i.e.

$$\underset{\approx}{d}\,\underset{\sim}{y} = \underset{\approx}{d}\,\underset{\approx}{X}_{(1)}\underset{\sim}{\theta}_{(1)} + \underset{\approx}{d}\,\underset{\sim}{f} \tag{4.33}$$



or

$$\underset{\sim}{y}_{(1)} = \underset{\approx}{W}\,\underset{\sim}{\theta}_{(1)} + \underset{\sim}{\dot{e}}_{(1)} \tag{4.34}$$

where

$$\underset{\sim}{y}_{(1)} = \begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = \underset{\approx}{d}\,\underset{\sim}{y} \; ; \quad \underset{\approx}{W} = \begin{bmatrix} \underset{\sim}{w}_2' \\ \vdots \\ \underset{\sim}{w}_n' \end{bmatrix} = \underset{\approx}{d}\,\underset{\approx}{X}_{(1)} \tag{4.35}$$

and

$$\begin{cases} \dot{e}_2 = M_1 + \dot{\varepsilon}_2 \\ \dot{e}_i = \phi_1 \dot{e}_{i-1} + \dot{\varepsilon}_i \qquad i = 3,4,\ldots,n \end{cases} \tag{4.36}$$

where the $\dot{\varepsilon}_i$'s are i.i.d $N(0,\sigma^2)$.

Hence the noise vector $\underset{\sim}{f}$ relating to the original observations can be represented by the stochastic model

$$(1-\phi_1 B)\,\nabla\, f_i = \dot{\varepsilon}_i \tag{4.37}$$

where $\nabla = (1-B)$. (4.37) is an ARIMA model of order (1,1,0), which incidentally in the present formulation also covers explosive situations, $1 < \phi_1$.

Introducing $H_0$ and $H_1$ for the two alternative models:

$$H_0 : \underset{\sim}{y} = \underset{\approx}{X}\,\underset{\sim}{\theta} + \underset{\sim}{e} \tag{4.38}$$

$$H_1 : \underset{\sim}{y} = \underset{\approx}{X}_{(1)}\underset{\sim}{\theta}_{(1)} + \underset{\sim}{f} \tag{4.39}$$

the distribution function of $\underset{\sim}{y}$ in relation to $H_0$ is given by Equation (4.6), since of course (4.38) is a repetition of (4.4).

In order to derive the distribution function of $\underset{\sim}{y}$ relative to $H_1$, we shall set $v_1 = y_1$ and define the parameter $\theta_0$ so that

$$\underset{\sim}{v} = \begin{bmatrix} v_1 \\ \underset{\sim}{y}_{(1)} \end{bmatrix} = \begin{bmatrix} \theta_0 \\ \underset{\approx}{W}\,\underset{\sim}{\theta}_{(1)} \end{bmatrix} + \begin{bmatrix} \dot{\varepsilon}_1 \\ \underset{\sim}{\dot{e}}_{(1)} \end{bmatrix} \tag{4.40}$$

where $\dot{\varepsilon}_1$ is distributed jointly with the $\dot{\varepsilon}_i$'s of (4.36) as

$$p(\underset{\sim}{\dot{\varepsilon}}) = (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}\sum_1^n \dot{\varepsilon}_i^2) \ . \tag{4.41}$$

From (4.41) and (4.36) we find

$$p(\underset{\sim}{\dot{e}}_{(1)}, \dot{\varepsilon}_1) = (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(\underset{\sim}{\dot{e}}'_{(1)} \underset{\approx}{C}_1 \underset{\sim}{\dot{e}}_{(1)} + (\dot{e}_2 - M_1)^2 + \dot{\varepsilon}_1^2)) \tag{4.42}$$

where

$$\underset{\approx}{C} = \underset{\approx}{C}'_1 \underset{\approx}{C}_1 \ ; \quad \underset{(n-2)\times(n-1)}{\underset{\approx}{C}_1} = \begin{bmatrix} -\phi_1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -\phi_1 & 1 \end{bmatrix} \tag{4.43}$$

Now (4.42) in conjunction with (4.40) yields

$$p(\underset{\sim}{y} | M_1, \underset{\sim}{\theta}_{(1)}, \theta_0, \phi_1, \sigma) \tag{4.44}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\tfrac{1}{2}\sigma^{-2}((\underset{\sim}{v}_{(1)} - \underset{\approx}{W}\underset{\sim}{\theta}_{(1)})' \underset{\approx}{C}_1 (\underset{\sim}{v}_{(1)} - \underset{\approx}{W}\underset{\sim}{\theta}_{(1)}) + (v_2 - \underset{\sim}{w}'_2 \underset{\sim}{\theta}_{(1)} - M_1)^2 + (v_1 - \theta_0)^2))$$

and since the transformation

$$\underset{n\times 1}{\underset{\sim}{y}} = \begin{bmatrix} 1 & & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \underset{n\times 1}{\underset{\sim}{y}} \tag{4.45}$$

has unit Jacobian, the right hand side of (4.44) is also the density function of $\underset{\sim}{y}$ :

$$p(\underset{\sim}{y} | M_1, \underset{\sim}{\theta}_{(1)}, \theta_0, \phi_1, \sigma) \tag{4.46}$$
$$= (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(SS_1(\underset{\sim}{\theta}_{(1)}, \phi_1) + (v_2 - \underset{\sim}{w}'_2 \underset{\sim}{\theta}_{(1)} - M_1)^2 + (v_1 - \theta_0)^2))$$

where

$$SS_1(\underset{\sim}{\theta}_{(1)}, \phi_1) = (\underset{\sim}{v}_{(1)} - \underset{\approx}{W}\underset{\sim}{\theta}_{(1)})' \underset{\approx}{C}_1 (\underset{\sim}{v}_{(1)} - \underset{\approx}{W}\underset{\sim}{\theta}_{(1)}) \tag{4.47}$$
$$= (\underset{\approx}{C}_1 \underset{\sim}{v}_{(1)} - \underset{\approx}{C}_1 \underset{\approx}{W}\underset{\sim}{\theta}_{(1)})' (\underset{\approx}{C}_1 \underset{\sim}{v}_{(1)} - \underset{\approx}{C}_1 \underset{\approx}{W}\underset{\sim}{\theta}_{(1)})$$



much like (4.9). When $\underset{\sim}{y}$ is considered known, (4.46) is of course the likelihood function for the parameters $(M_1, \underset{\sim}{\theta}_{(1)}, \theta_0, \phi_1, \sigma)$, and it is maximized by minimizing the (residual) sum of squares of the estimated $\dot{\varepsilon}_i$'s, $SS_1(\underset{\sim}{\theta}_{(1)}, \phi_1)$; the minimum is $SS_1(\hat{\underset{\sim}{\theta}}_{(1)}, \hat{\phi}_1)$.

Since $H_0$ and $H_1$ have the same number of parameters, $p + 3$, we may formally equate

$$\begin{cases} p(\underset{\sim}{y} | \underset{\sim}{\psi}, H_0) = p(\underset{\sim}{y} | M_0, \underset{\sim}{\theta}, \phi_0, \sigma) \\ p(\underset{\sim}{y} | \underset{\sim}{\psi}, H_1) = p(\underset{\sim}{y} | M_1, \underset{\sim}{\theta}_{(1)}, \theta_0, \phi_1, \sigma) \end{cases} \tag{4.48}$$

and these distributions are known exactly including their normalizing constants.

### 4.3.2 The complementing priors.

Before posterior model probabilities can be computed, prior distributions for $\underset{\sim}{\psi}$ complementing the likelihood functions must be produced.

As far as $\ell_0 \propto p(\underset{\sim}{y} | \underset{\sim}{\psi}, H_0)$ is concerned, we have (4.10):

$$p(\underset{\sim}{\psi} | H_0) = p(M_0, \underset{\sim}{\theta}_1, \phi_0, \sigma) = |\underset{\approx}{X}' \underset{\approx}{C}_0 \underset{\approx}{X}|^{1/2} \sigma^{-1} p(\phi_0) \tag{4.49}$$

An approximately noninformative prior complementing the nonstationary alternative likelihood function function $\ell_1 \propto p(\underset{\sim}{y} | \underset{\sim}{\psi}, H_1)$ is derived as follows. On an assumption of independence a priori between the linear parameter $\theta_0$ measuring an initial level of the series, and the remaining parameters, we may write

$$p(\underset{\sim}{\psi} | H_1) = p(M_1, \underset{\sim}{\theta}_{(1)}, \theta_0, \phi_1, \sigma) = p(M_1, \underset{\sim}{\theta}_{(1)}, \phi_1, \sigma) p(\theta_0) \tag{4.50}$$

where now the prior $p(M_1, \underset{\sim}{\theta}_{(1)}, \phi_1, \sigma)$ may be viewed as complementing the conditional likelihood

$$\ell_1(M_1,\underset{\sim}{\theta}_{(1)},\phi_1,\sigma|\theta_0,\underset{\sim}{y}) = \ell_1(M_1,\underset{\sim}{\theta}_{(1)},\phi_1,\sigma|\underset{\sim}{v}_{(1)})$$

$$\propto p(\underset{\sim}{v}_{(1)}|M_1,\underset{\sim}{\theta}_{(1)},\phi_1,\sigma)$$

$$= (2\pi\sigma^2)^{-\frac{n-1}{2}} \exp\left(-\tfrac{1}{2}\sigma^{-2}\left((\underset{\sim}{v}_{(1)}-\underset{\approx}{W}\underset{\sim}{\theta}_{(1)})'\underset{\approx}{C}_1(\underset{\sim}{v}_{(1)}-\underset{\approx}{W}\underset{\sim}{\theta}_{(1)})+(v_2-\underset{\sim}{w}_2'\underset{\sim}{\theta}_{(1)}-M_1)^2\right)\right) \tag{4.51}$$

which has the familiar AR-1 form as before, (4.6). Hence on the same arguments (presented in Chapter 2) that led to the adoption of the approximately noninformative prior (4.49) in conjunction with $H_0$, and assigning a locally uniform prior for the location parameter $\theta_0$, we find

$$p(\underset{\sim}{\psi}|H_1) \propto |\underset{\approx}{W}'\underset{\approx}{C}_1\underset{\approx}{W}|^{1/2}\,\sigma^{-1}\,p(\phi_1) \tag{4.52}$$

where as before $p(\phi_1)$ is a locally approximately noninformative prior for $\phi_1$.

*Equations (4.49) and (4.52) specify the priors* $p(\underset{\sim}{\psi}|H_m)$, $m = 0,1$, up to a multiplicative constant only. In order to compute posterior model probabilities, it is necessary to know the ratio of these proportionality constants; in fact prior model probabilities need also be stated. First utilizing the (usual) assumption of prior independence between $\sigma$ and the remaining parameters we may factorize:

$$p(\underset{\sim}{\psi}|H_0) = p(M_0,\underset{\sim}{\theta},\phi_0,\sigma) = p(\underset{\sim}{\theta}|M_0,\phi_0)\,p(M_0|\phi_0)p(\phi_0)p(\sigma) \tag{4.53}$$

and

$$p(\underset{\sim}{\psi}|H_1) = p(M_1,\underset{\sim}{\theta}_{(1)},\theta_0,\phi_1,\sigma) = p(\underset{\sim}{\theta}_{(1)},\theta_0|M_1,\phi_1)p(M_1|\phi_1)p(\phi_1)p(\sigma). \tag{4.54}$$

Now further progress can be made, if we make the following assumptions:



1) The constants of proportionality in

$$p(\underset{\sim}{\theta}|M_0,\phi_0) \propto |\underset{\approx}{X}'\underset{\approx}{C}_0\underset{\approx}{X}|^{1/2} \tag{4.55}$$

and

$$p(\underset{\sim}{\theta}_{(1)},\theta_0|M_1,\phi_1) \propto |\underset{\approx}{W}'\underset{\approx}{C}_1\underset{\approx}{W}|^{1/2} \tag{4.56}$$

are the same.

It is noted, that conditionally on $M_m,\phi_m$ $m = 0,1$, the two alternative models are equivalent to

$$H_0|M,\phi: \qquad \underset{\sim}{z}_0 = \underset{\approx}{\Xi}_0\underset{\sim}{\theta} + \underset{\sim}{\varepsilon} \tag{4.57}$$

with $\underset{\sim}{z}_0 = \underset{\approx}{C}_0\underset{\sim}{Y}$ ; $\underset{\approx}{\Xi}_0 = \underset{\approx}{C}_0\underset{\approx}{X}$

$$H_1|M,\phi: \qquad \underset{\sim}{z}_1 = \underset{\sim}{\Xi}_1\begin{pmatrix}\theta_0\\ \underset{\sim}{\theta}_{(1)}\end{pmatrix} + \underset{\sim}{\dot{\varepsilon}} \tag{4.58}$$

with $\underset{\sim}{z}_1 = \begin{bmatrix} v_1 \\ \underset{\approx}{C}_1\underset{\sim}{v}_{(1)}\end{bmatrix}$; $\underset{\approx}{\Xi}_1 = \begin{bmatrix} 1 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \underset{\approx}{C}_1\underset{\approx}{W}\end{bmatrix}$

and models (4.57) and (4.58) are ordinary linear models of p linear parameters, and i.i.d. $N(0,\sigma^2)$ noise.

Now if we employ to this situation the argument of Kanemasu [1973], that the increase in information (measured by entropy) about the parameters $\underset{\sim}{\theta}$ and $(\underset{\sim}{\theta}_{(1)},\theta_0)$ from n observations, is expected to be the same a priori; then his result says, that indeed

$$\frac{p(\underset{\sim}{\theta}|M_0,\phi_0)}{p(\underset{\sim}{\theta}_{(1)},\theta_0|M_1,\phi_1)} = \frac{|\underset{\approx}{\Xi}_0'\underset{\approx}{\Xi}_0|^{1/2}}{|\underset{\approx}{\Xi}_1'\underset{\approx}{\Xi}_1|^{1/2}} = \frac{|\underset{\approx}{X}'\underset{\approx}{C}_0\underset{\approx}{X}|^{1/2}}{|\underset{\approx}{W}'\underset{\approx}{C}_1\underset{\approx}{W}|^{1/2}} \tag{4.59}$$

2) Since $\sigma$ in the two models is the same parameter, the proportionality factor is automatically the same in both cases.

3) $p(\phi_0)$ and $p(\phi_1)$ being noninformative priors for $\phi_0$ and $\phi_1$ respectively may equivalently be expressed as locally uniform priors in an approximately data translating metric (see Appendix A of

Chapter 2), i.e.

$$\begin{cases} p(\phi_0)\, d\phi_0 = k_0\, d\phi_0 \\ p(\phi_1)\, d\phi_1 = k_1\, d\phi_1 \end{cases} \quad . \tag{4.60}$$

Since $\phi_0$ and $\phi_1$ are parameters of the same kind, we shall assume $k_0 = k_1$.

4) The locally uniform priors for the linear starting parameters $M_0$ and $M_1$ are similarly assumed to have the same constant of proportionality, as they too are parameters of the same kind

$$p(M_0|\phi_0) = p(M_1|\phi_1) \tag{4.61}$$

5) Finally prior indifference between $H_0$ and $H_1$ is expressed as

$$P(H_0) = P(H_1) \quad . \tag{4.62}$$

4.3.3 Posterior model probabilities.

In general, if a number of alternative models $H_m$, $m = 1,2,\ldots,N_H$, were under consideration, and it was known, that the data generating process was actually counted among these models, then Bayes' theorem states that

$$P(H_m|\underset{\sim}{y}) = \frac{p(\underset{\sim}{y}|H_m)\, P(H_m)}{\sum_m p(\underset{\sim}{y}|H_m) P(H_m)} \tag{4.63}$$

Because such knowledge is unavailable in practice, or more fundamentally, because no model is ever exactly correct, probability statements computed from (4.63) cannot be given an absolute interpretation. Rather they are quantitative expressions of the evidence in the data concerning the relative appropriateness of the suggested alternatives. So that even if perhaps some (unknown) model $H_a$ might be superior to $H_0$ and $H_1$, and in a three way comparison would have claimed most of the posterior probatility, it is still legimate to write:



$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \frac{p(\underset{\sim}{y}|H_0)\, P(H_0)}{p(\underset{\sim}{y}|H_1)\, P(H_1)} = \frac{\int_{\underset{\sim}{\psi}} p(\underset{\sim}{y}|\underset{\sim}{\psi},H_0)\, p(\underset{\sim}{\psi}|H_0)\, P(H_0)\, d\underset{\sim}{\psi}}{\int_{\underset{\sim}{\psi}} p(\underset{\sim}{y}|\underset{\sim}{\psi},H_1)\, p(\underset{\sim}{\psi}|H_1)\, P(H_1)\, d\underset{\sim}{\psi}}$$

$$= \frac{P(H_0)}{P(H_1)} \frac{\int_{M_0,\underset{\sim}{\theta},\phi_0,\sigma} I_{H_0}(M_0,\underset{\sim}{\theta},\phi_0,\sigma)\, dM_0 d\underset{\sim}{\theta} d\phi_1 d\sigma}{\int_{M_1,\underset{\sim}{\theta}_{(1)},\theta_0,\phi_1,\sigma} I_{H_1}(M_1,\underset{\sim}{\theta}_{(1)},\theta_0,\phi_1,\sigma)\, dM_1 d\underset{\sim}{\theta}_{(1)} d\theta_0 d\phi_1 d\sigma}$$

$$= \frac{\int_{\phi_0} p(\phi_0)\, (SS_0(\hat{\underset{\sim}{\theta}},\phi_0))^{-\frac{n-p-1}{2}}\, d\phi_0}{\int_{\phi_1} p(\phi_1)\, (SS_1(\hat{\underset{\sim}{\theta}}_{(1)},\phi_1))^{\frac{n-p-1}{2}}\, d\phi_1} \tag{4.64}$$

with

$$I_{H_0}(M_0,\underset{\sim}{\theta},\phi_0,\sigma) =$$

$$(2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(SS_0(\underset{\sim}{\theta},\phi_0)+(e_1-M_0)^2))\, |\underset{\approx}{X}'\underset{\approx}{C}_0\underset{\approx}{X}|^{1/2}\, \sigma^{-1} p(\phi_0)$$

and

$$I_{H_1}(M_1,\underset{\sim}{\theta}_{(1)},\theta_0,\phi_1,\sigma) =$$

$$(2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2}\sigma^{-2}(SS_1(\underset{\sim}{\theta}_{(1)},\phi_1)+(\underset{\sim}{e}_2-M_1)^2+(v_1-\theta_0)^2))\, |\underset{\approx}{W}'\underset{\approx}{C}_1\underset{\approx}{W}|^{1/2}\, \sigma^{-1} p(\phi_1).$$

Unlike the integrations leading to the last expression in (4.64), the still remaining integration over $\phi$ cannot be carried out analytically. However in terms of the data translating matrix $\Phi$ induced by $p(\phi)$, we may write

$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \frac{(SS_0(\hat{\underset{\sim}{\theta}},\hat{\phi}_0))^{-\frac{n-p-1}{2}} \int_{\phi_0} (SS_0(\hat{\underset{\sim}{\theta}},\phi_0)/SS_0(\hat{\underset{\sim}{\theta}},\hat{\phi}_0))^{-\frac{n-p-1}{2}}\, d\phi_0}{(SS_1(\hat{\underset{\sim}{\theta}}_{(1)},\hat{\phi}_1))^{-\frac{n-p-1}{2}} \int_{\phi_1} (SS_1(\hat{\underset{\sim}{\theta}}_{(1)},\phi_1)/SS_1(\hat{\underset{\sim}{\theta}}_{(1)},\hat{\phi}_1))^{\frac{n-p-1}{2}}\, d\phi_1} \tag{4.65}$$

and having now scaled the integrand so that its maximum value is one, it is recognized, that, by virtue of the principle on which the approximate data translating metric $\Phi$ rests (i.e. constant expected spread) the integrals in (4.65) are nearly constant, not depending on the data (see Appendix A of Chapter 2).

Hence we have established the remarkably simple

Rule: The evidence in the data, $\underset{\sim}{y}$, concerning whether serial correlation is better accounted for by an AR-1 noise model ($H_0$) with a fixed mean, or by a nonstationary noise model ($H_1$) implying that the differenced data have AR-1 noise, is expressed through the minimum residual sum of squares of each model; specifically

$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \left( \frac{SS_0(\hat{\underset{\sim}{\theta}}, \hat{\phi}_0)}{SS_1(\hat{\underset{\sim}{\theta}}_{(1)}, \hat{\phi}_1)} \right)^{-\frac{n-p-1}{2}} . \qquad (4.66)$$

If one had been unwilling to make the assumptions 1) through 5) in Subsection 4.3.2, then Equation (4.66) would have been derived up to a multiplicative constant $\kappa$. In favor of those assumptions is the fact, that they lead to $\kappa = 1$, an intuitively very appealing result.

## 4.4 The CGK data.

In this reexamination of the CGK data, we shall use all 55 observations actually listed in their paper[†].

Analyzing these data on the basis of model $H_0$ (4.38) the minimum residual sum of squares is $SS_0(\hat{\underset{\sim}{\theta}}, \hat{\phi}_0 = .80) = 15624$.

The marginal posterior distribution of $\phi_0$, $p(\phi_0|\underset{\sim}{y})$ (4.10) is shown in Figure 4.1a. It rather conclusively rules out the possibility

[†] CGK only used the first 51 observations in their analysis, and so did B&N in their reanalysis.



of $\phi_0$ being zero; the body of the distribution lies in the positive stationary range of $\phi_0$.

Posterior distributions of $\theta_1$ $\theta_2$ and $\theta_3$ are given in Figures 4.2a,b and c respectively. The curves $p_0(\theta_j|\phi_0 = 0, \underset{\sim}{y})$ show the results of relying on an assumption of independence, i.e. they correspond to the CGK analysis. The curves $p_0(\theta_j|\phi_0 = \hat{\phi}_0, \underset{\sim}{y})$ correspond to the analysis of B&N, while the marginal distributions $p_0(\theta_j|\underset{\sim}{y})$ represent the Bayesian inference. It is seen, that the marginal posterior distributions and the posterior distribution conditional on $\phi_0 = \hat{\phi}_0$ are almost identical for $\theta_1$ and $\theta_2$ and indistinguishable for $\theta_3$. On the other hand the conditional distributions $p_0(\theta_j|\phi_0 = 0, \underset{\sim}{y})$ differ considerably from those, both in spread and location.

The cause of these shifts are made clear by looking at the joint posterior distribution (4.16) of $\{\theta_j, \phi_0\}$ $j = 1,2,3$ shown in Figures 4.3a,b and c. The contours of the HPD regions with probability contents 50%, 90%, 95% and 99% were computed using numerical integration, while those labeled 99.9%, 99.99%, 99.999% and 99.9999% were constructed using (4.18); and they are to be considered rough guides only since the normal approximation may be poor.

These contour plots show, that the parameters $\theta_j$ and $\phi_0$ are dependent. It is seen, that the distribution of $\theta_j$ conditional on $\phi_0$ shifts location as well as change in spread as $\phi_0$ moves along its axis. In particular if $\phi_0$ is set equal to the extremely unlikely value zero, then the shift is quite large. The 95% HPD intervals for $\theta_j$, $j = 1,2,3$, are traced as a function of $\phi_0$ in these contour plots Figures 4.3a,b and c (as well as in the plots of Figures 4.3d, 4.6a,

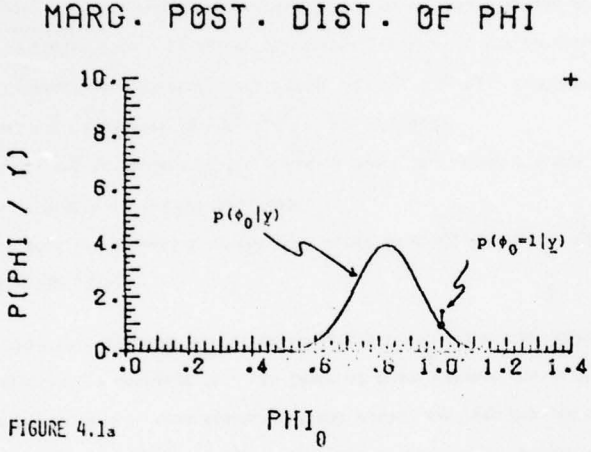



FIGURE 4.1a

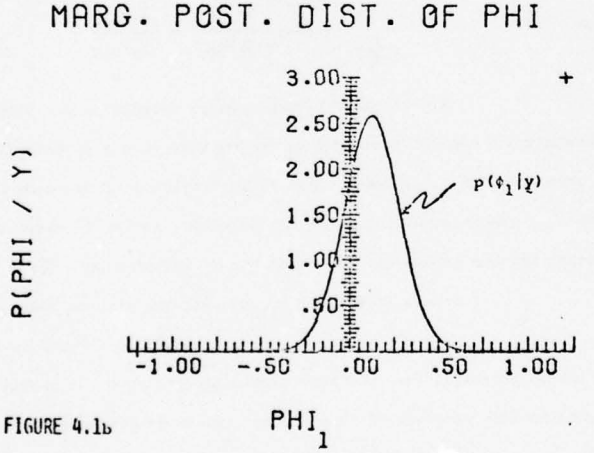



FIGURE 4.1b



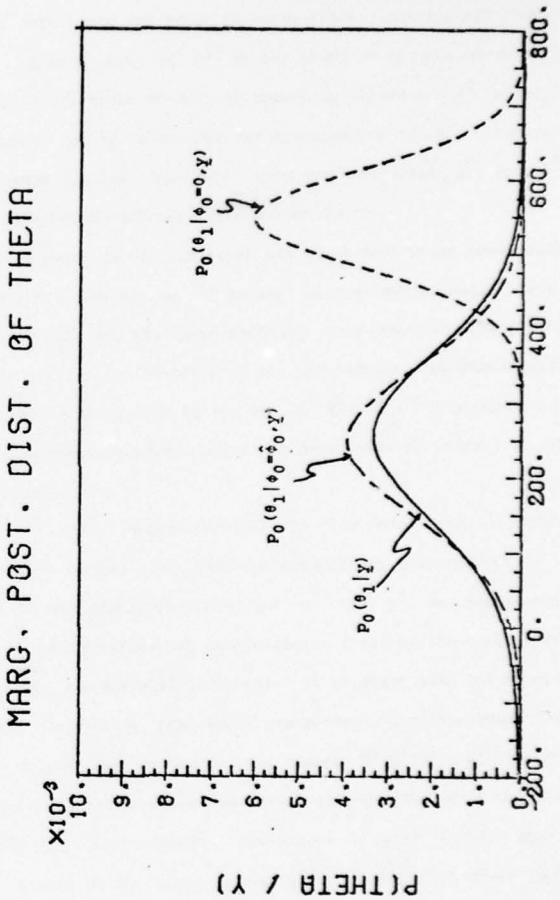



FIGURE 4.2a

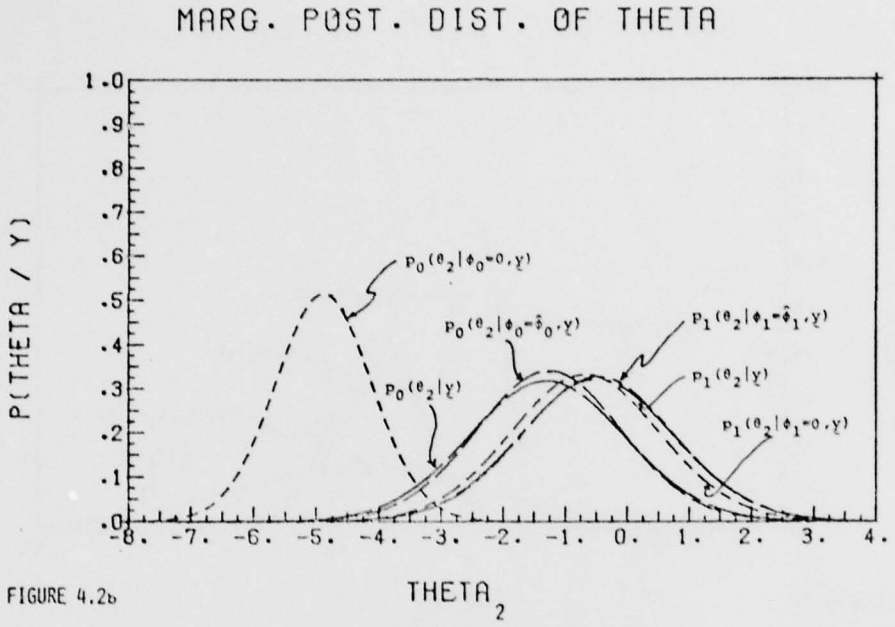
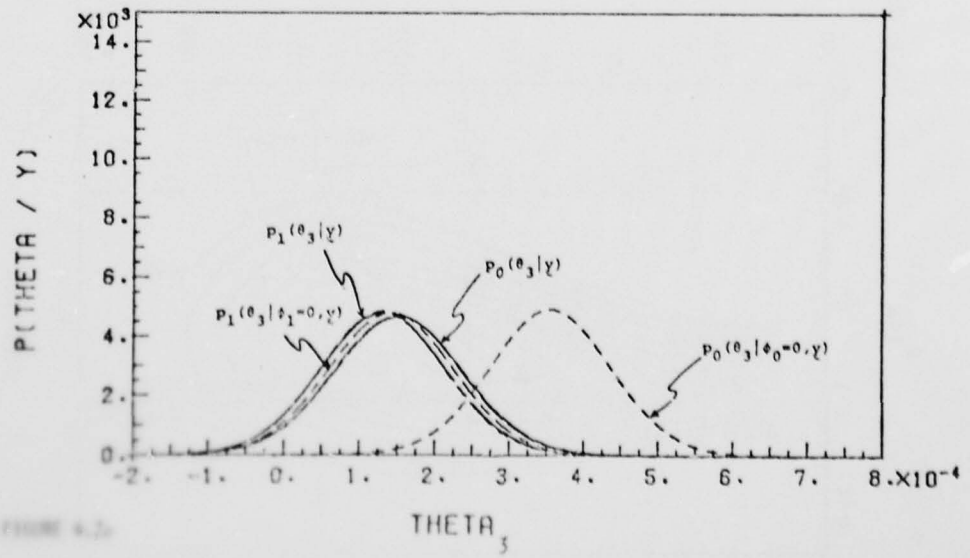

MARG. POST. DIST. OF THETA
P(THETA / Y)
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.0
$P_0(\theta_2|\phi_0=0,\underline{Y})$
$P_0(\theta_2|\phi_0=\hat{\phi}_0,\underline{Y})$
$P_1(\theta_2|\phi_1=\hat{\phi}_1,\underline{Y})$
$P_0(\theta_2|\underline{Y})$
$P_1(\theta_2|\underline{Y})$
$P_1(\theta_2|\phi_1=0,\underline{Y})$
-8. -7. -6. -5. -4. -3. -2. -1. 0. 1. 2. 3. 4.
$THETA_2$

FIGURE 4.2b




MARG. POST. DIST. OF THETA
P(THETA / Y)
$\times 10^3$
14.
12.
10.
8.
6.
4.
2.
0.
$P_1(\theta_3|\underline{Y})$
$P_0(\theta_3|\underline{Y})$
$P_1(\theta_3|\phi_1=0,\underline{Y})$
$P_0(\theta_3|\phi_0=0,\underline{Y})$
-2. -1. 0. 1. 2. 3. 4. 5. 6. 7. 8. $\times 10^{-4}$
$THETA_3$

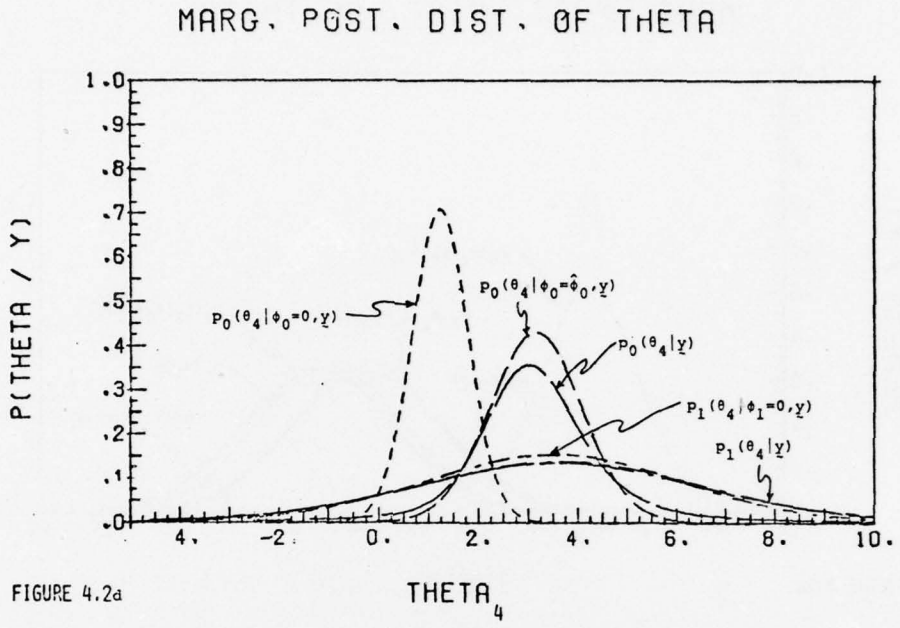



FIGURE 4.2a



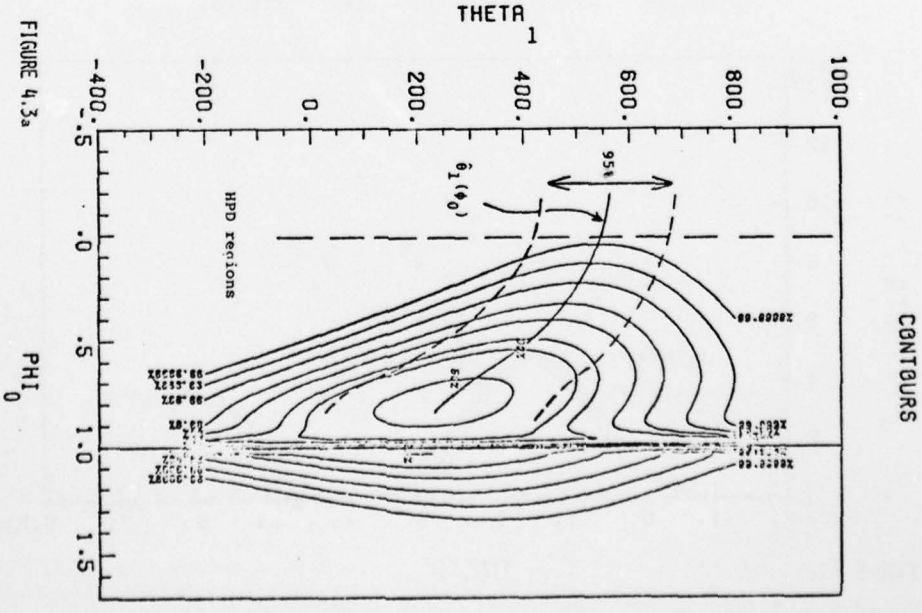



FIGURE 4.3a

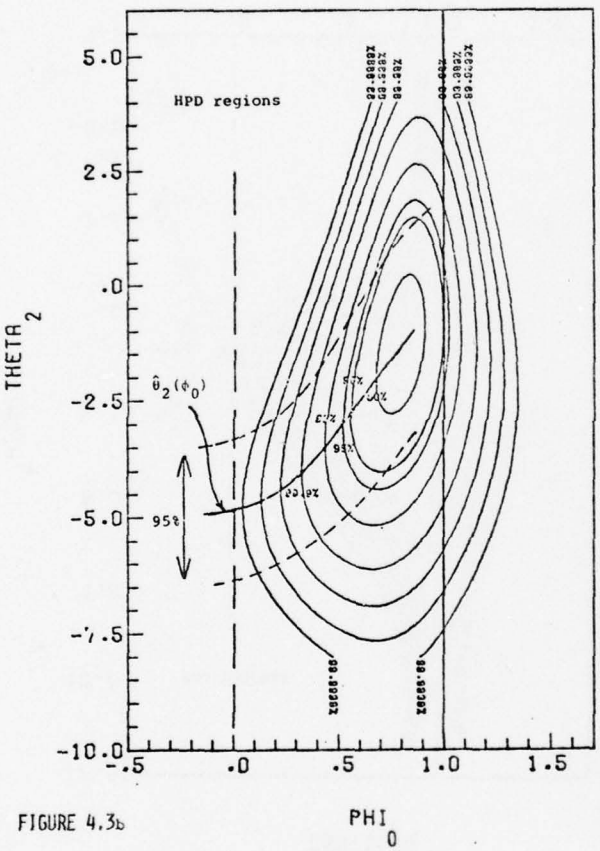



FIGURE 4.3b



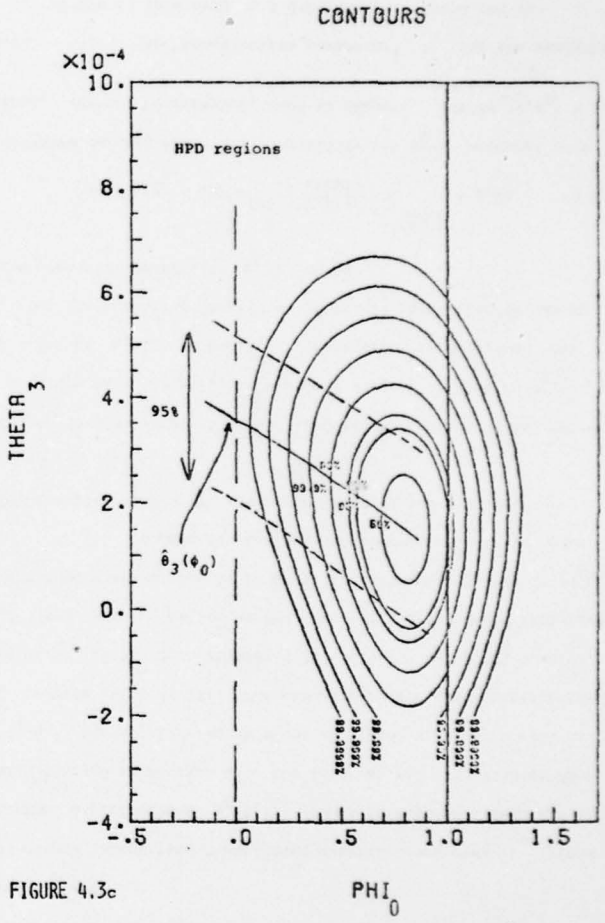



FIGURE 4.3c

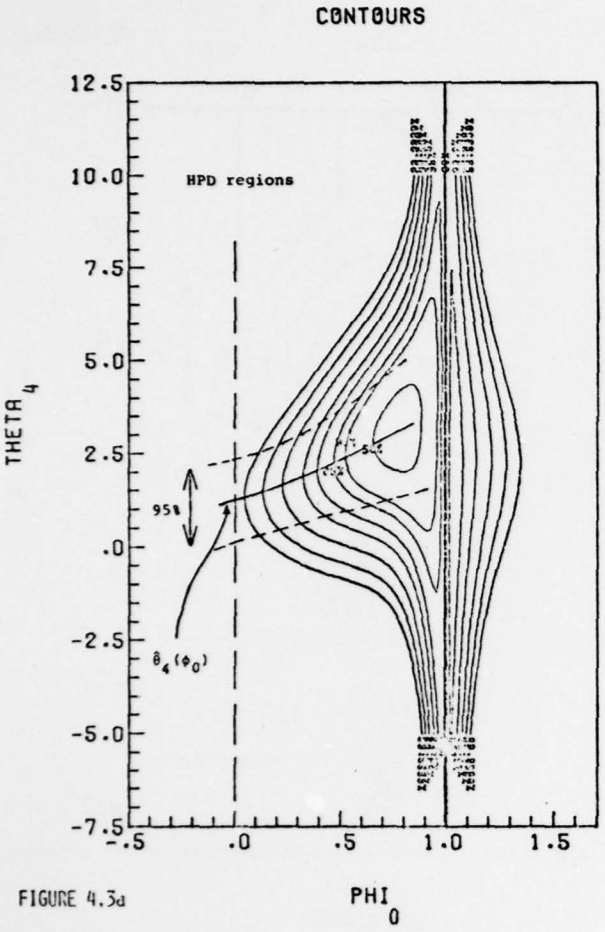



FIGURE 4.3a



b and c). It is reminded, that the 95% HPD interval for $\theta_j$ conditional on $\phi_0 = \phi_0'$ is identical to the 95% confidence interval for $\theta_j$ which would have been derived from a sampling theory point of view assuming $\phi_0 = \phi_0'$.

Of course integrating these joint distributions over $\phi_0$ yield the marginal distributions $p_0(\theta_j|\underset{\sim}{y})$ shown in Figures 4.2a,b and c.

As discussed in Section 4.1, the line $\phi_0 = 1$ is excluded from the $\{\theta_j,\phi_0\}$ plane. Inferences about $\theta_2$ and $\theta_3$ conditional on $\phi_0 = 1$ is made using (4.23). The conditional posterior distribution $p_0(\theta_j|\phi_0=1,\underset{\sim}{y})$ is found in Figures 4.2b and 4.2c as $p_1(\theta_j|\phi_1=0,\underset{\sim}{y})$, $j = 1,2$ (see below). As far as the mean $\theta_1$ is concerned, this parameter vanishes when $\phi_0=1$. It is seen from Figure 4.3a that as $\phi_0$ approaches 1, $\theta_1$ becomes increasingly indeterminate. It is noted, that for example the 95% HPD region consists of two "islands" on either side of $\phi_0=1$.

To see to what extent the data supports the notion of the noise being a random walk ($\phi_0=1$), one may more closely examine $p(\phi_0|\underset{\sim}{y})$ Figure 4.1a at $\phi_0=1$. If continuity were maintained through this point, then its posterior density would be .92, but using (4.32) we find the posterior density at $\phi_0=1$ to be

$$p(\phi_0=1|\underset{\sim}{y}) = .92\, e^{1/2}\left(\frac{17179.}{17152.}\right)^{-\frac{55-4-1}{2}} = 1.46 \qquad (4.67)$$

This elevated density makes the possibility of $\phi_0=1$ somewhat more plausible. (It may be remarked, that to compute $\lim_{\phi_0 \to 1} SS_0(\hat{\underset{\sim}{\theta}},\phi_0) = SS_0(\hat{\underset{\sim}{\theta}},\phi_0) = 17152$. the "substitution parameter" $\theta_1^*$ for the vanishing mean $\theta_1$, serves in this case as a deterministic trend for the

differences and not as mean for the differences, since $\theta_4$ assume that role when $\phi_0=1$; see Section 3.2 of Chapter 3).

If alternatively the differenced data are analyzed along the lines of (4.34), i.e. if $\underset{\sim}{y}$ is considered generated by $H_1$ (4.39), then the residual sum of squares is $SS_1(\hat{\theta}_{(1)}, \hat{\phi}_1 = .12) = 16853.$, i.e. not quite as good a fit as $H_0$ provided.

Figure 4.1b shows, that the autoregressive parameter $\phi_1$ for the differences has it marginal posterior distribution located near $\phi_1=0$.

The marginal posterior distributions $p_1(\theta_j|\underset{\sim}{y})$, $j = 2,3$, are shown in Figures ..2b and 4.2c. $p_1(\theta_j|\phi_1=\hat{\phi}_1,\underset{\sim}{y})$ $j = 2,3$ are also drawn[†], these distributions are virtually indistinguishable from the marginal distributions. Also the posterior $p_1(\theta_j|\phi_1=0,\underset{\sim}{y})$ are drawn, since $\phi_1 = 0$ means that the noise $\underset{\sim}{f}$ is a random walk, these curves may also be interpreted as $p_0(\theta_j|\phi_0=1,\underset{\sim}{y})$ (except one observation has been lost in the beginning of the series).

Turning to what inferences can be drawn about the trend $\theta_4$, we observe in Figure 4.2d, that differencing has a marked effect on what the data have to say about $\theta_4$. It is seen, that if model $H_1$ is used (or if $\phi_0 = 1$ in $H_0$), then the need for having a trend in the model becomes much less persuasive. The question now arises of how much the testimony of $H_1$ shall be discounted relative to that of $H_0$, due to $H_1$'s slightly inferior fit. Putting the rule (4.66) to use we find

[†] These conditional distributions correspond essentially to the inference made by B&N in connection with their IMA-1 noise model, as that model is very nearly the same as the present ARIMA (1,1,0) for small $\phi_1$.



$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \left(\frac{SS_0}{SS_1}\right)^{-\frac{55-4-1}{2}} = 6.64 \qquad (4.68)$$

or

$$\begin{cases} P(H_0|\underset{\sim}{y}) = .87 \\ P(H_1|\underset{\sim}{y}) = .13 \end{cases} \qquad (4.69)$$

So as far as the data are concerned it would be a mistake to completely disregard the findings associated with $H_1$, in particular it may be contemplated to reanalyze the data without having a deterministic trend in the model. Before doing that we shall mention, that in situations like the present one, where the posterior model probabilities are not very conclusive, the possibility exists of combining the marginal posterior densities of the common linear parameters $\theta_j$, $j = 2,3,4$ from the two models. Conditional on $H_0$ or $H_1$ being an adequate model (i.e. $P(H_0) + P(H_1) = 1$) we may write

$$p(\theta_j|\underset{\sim}{y}) = p_0(\theta_j|H_0,\underset{\sim}{y})P(H_0|\underset{\sim}{y}) + p_1(\theta_j|H_1,\underset{\sim}{y})P(H_1|\underset{\sim}{y}) \qquad (4.70)$$

in other words a choice between $H_0$ and $H_1$ is not really necessary in order to make inferences about $\theta_j$, $j = 2,3,4$.

Reanalyzing the data without a deterministic trend in the model, i.e. rewriting (4.1) as

$$y_i = \theta_1 + \theta_2 x_{1,i} + \theta_3 x_{2,i} + N_i \qquad (4.71)$$

we find for $H_0$, that the marginal posterior distribution of $\phi_0$, Figure 4.4a, has shifted dramatically towards larger values (around 1), whereas $p(\phi_1|\underset{\sim}{y})$ of model $H_1$ has hardly changed at all, Figure 4.4b.

Looking at the density $p(\phi_0|\underset{\sim}{y})$ at $\phi_0=1$, we see that if continuity were maintained then this point would have a density of about 6.9; but using (4.32) we find

$$p(\phi_0=1|\underset{\sim}{y}) = 6.9\, e^{1/2} \left(\frac{17895.}{17179.}\right)^{-\frac{55-3-1}{2}} = 4.0 \,. \qquad (4.72)$$

Hence the inducement to difference has increased by the elimination of the trend from the model, but the increase is not as pronounced as it would have appeared, if $p(\phi_0|\underset{\sim}{y})$ had been interpreted as going continuously through $\phi_0=1$.

Applying the rule (4.66) we also find, that upon elimination of $\theta_4$, the model structure $H_1$ gains support vis-a-vis $H_0$. Specifically

$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \left(\frac{SS_0(\hat{\underset{\sim}{\theta}}, \hat{\phi}_0 = .96)}{SS_1(\hat{\underset{\sim}{\theta}}_{(1)}, \hat{\phi}_1 = .15)}\right)^{-\frac{55-3-1}{2}} = \left(\frac{16911.}{17379.}\right)^{-\frac{51}{2}} = 2.00 \qquad (4.73)$$

or

$$\begin{cases} P(H_0|\underset{\sim}{y}) = .67 \\ P(H_1|\underset{\sim}{y}) = .33 \,. \end{cases} \qquad (4.74)$$

For illustration Figures 4.6a, b and c show the joint posterior distribution of $\{\theta_j, \phi_0\}$ $j = 1,2,3$ in relation to model $H_0$. Now as before these contour plots demonstrate the dependence of $\theta_j$ on $\phi_0$. Integrating the joint distributions over $\phi_0$ yields the marginal posteriors $p_0(\theta_j|\underset{\sim}{y})$ $j = 1,2,3$ drawn in Figures 4.5a, b and c. These figures also show the posterior distributions conditional on $\phi_0=0$ and $\phi_0 = \hat{\phi}_0$, as well as the marginal and conditional distributions of $\theta_j$, $j = 2,3$ relating to model $H_1$. It is noted, that whatever disagreement existed between $p_0(\theta_j|\underset{\sim}{y})$ vs. $p_1(\theta_j|\underset{\sim}{y})$, $j = 2,3$, has now disappeared.

The marginal posterior distributions of $\theta_2$ and $\theta_3$ concur with the conclusion drawn by B&N, that "there is no real evidence of any



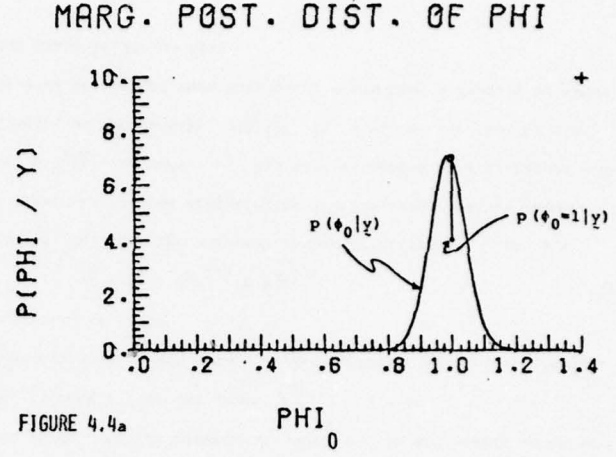



FIGURE 4.4a

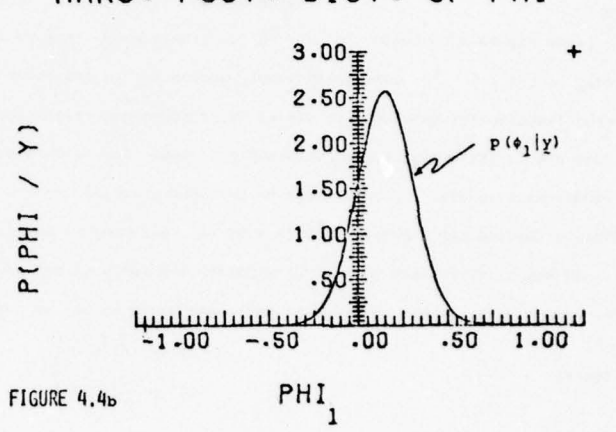



FIGURE 4.4b

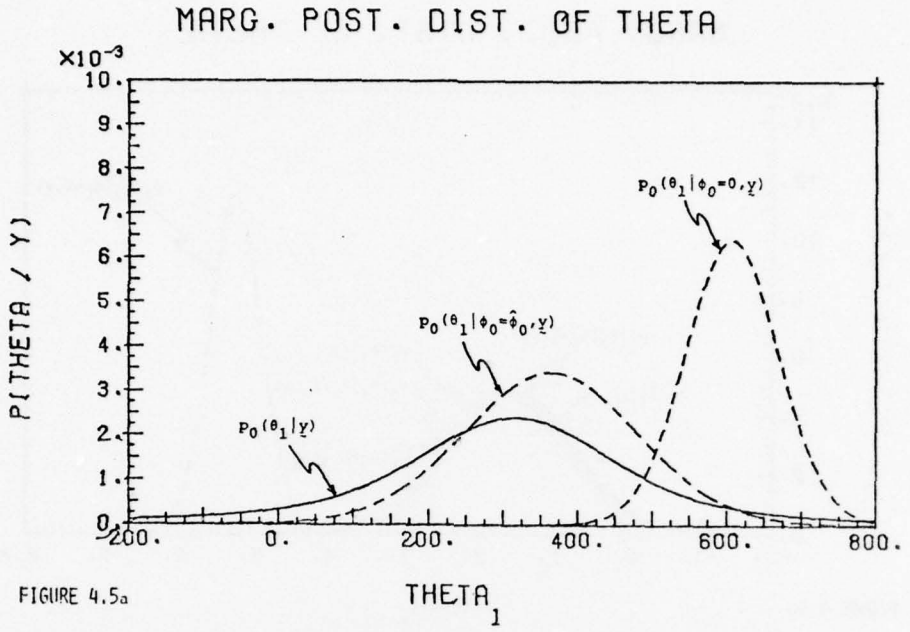



FIGURE 4.5a



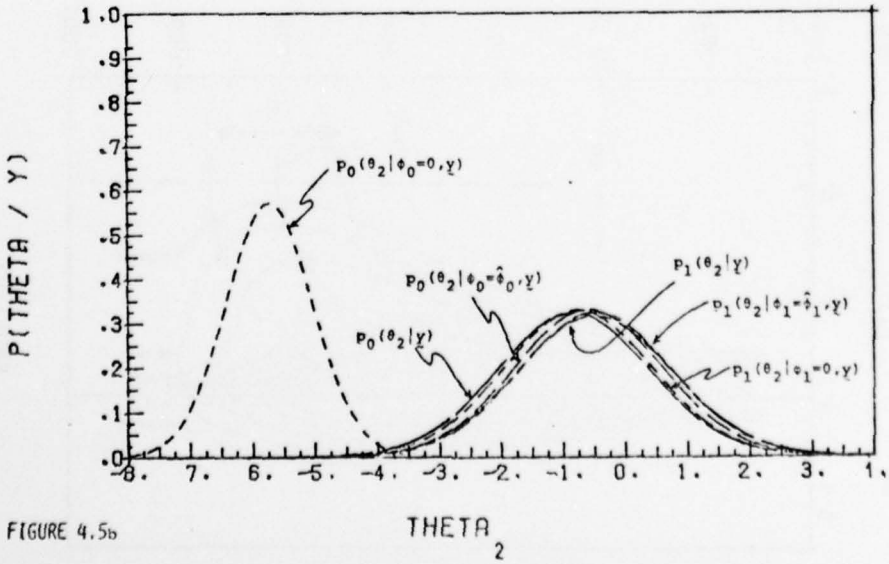



FIGURE 4.5b

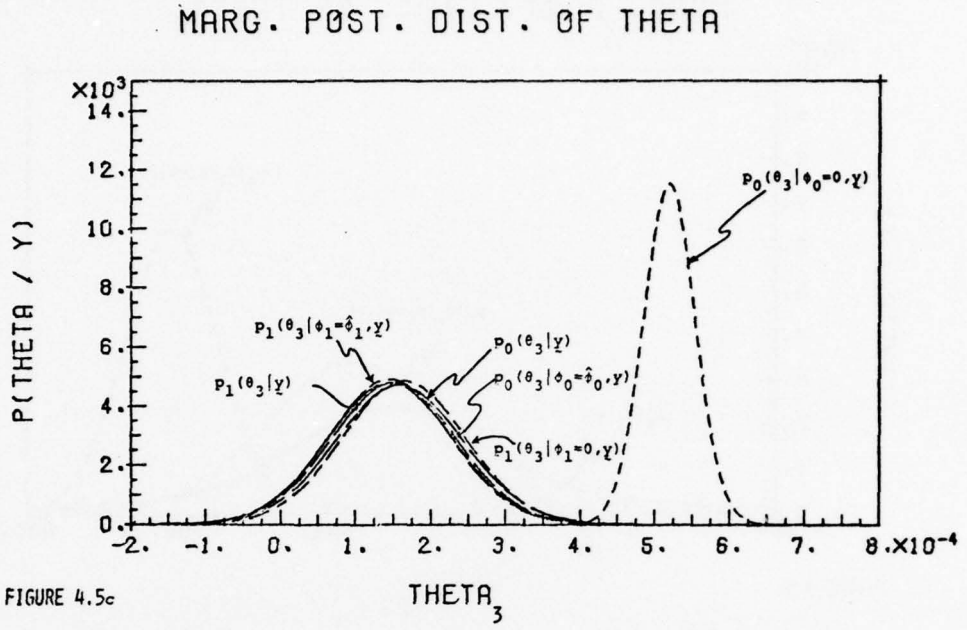



FIGURE 4.5c



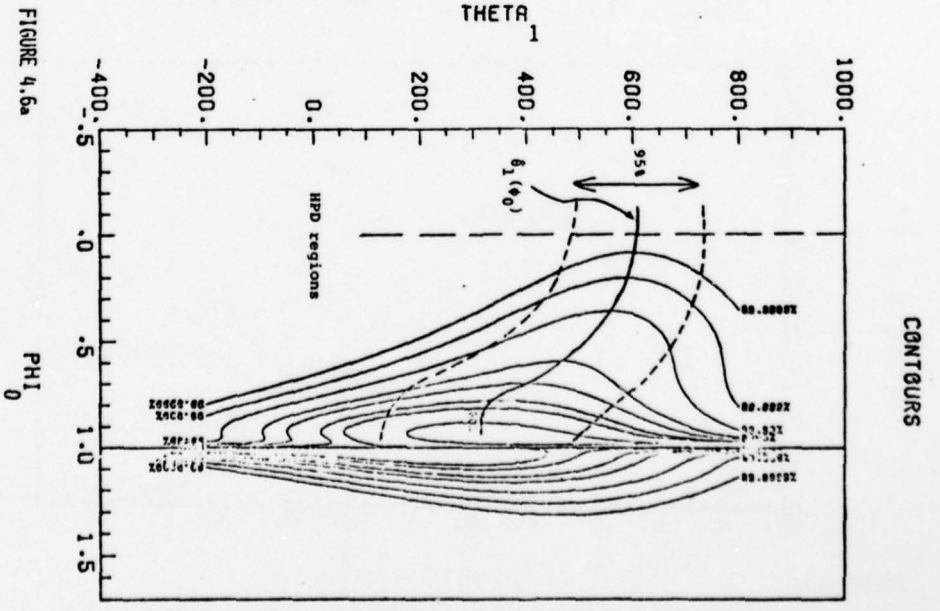



FIGURE 4.6a

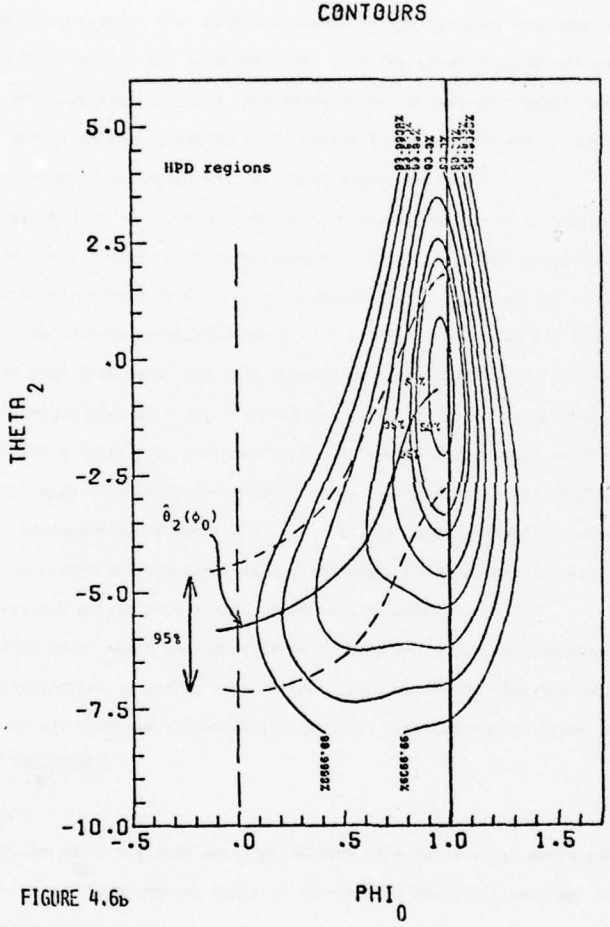



FIGURE 4.6b



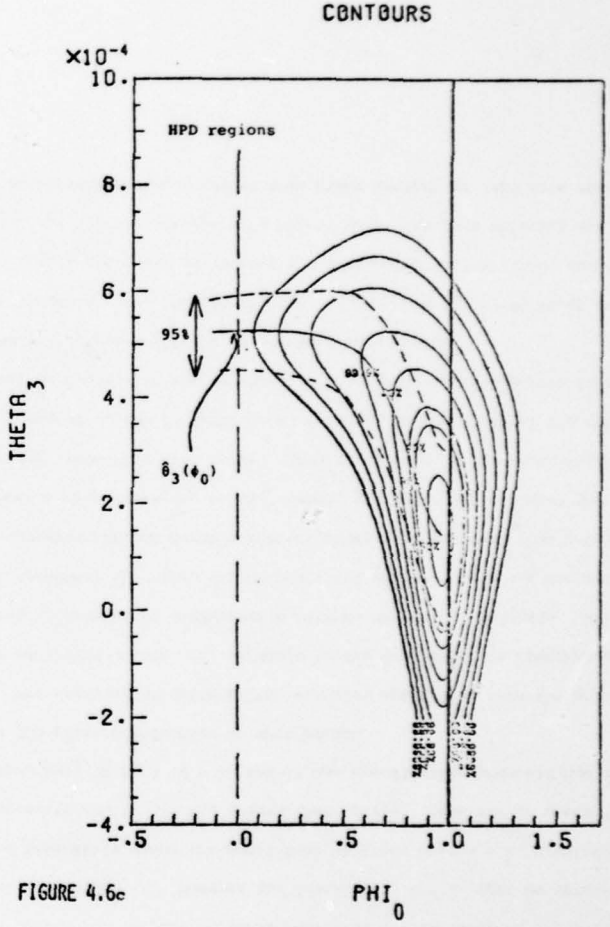



FIGURE 4.6c

relation between the output on the one hand and the two lagged input on the other". The tail area for $p(\theta_2|\underset{\sim}{y})$ to the right of $\theta_2 = 0$ is about 28%, and the tail area to the left of $\theta_3 = 0$ is about 4%; which is not particularly striking in light of the fact, that the proposed variables and lags were the results of a large scale selection effort.

4.5 Conclusion.

In the analysis of sequential data, it is necessary to allow for the possibility of serial correlation. For example in economic and business data, where observations have to be taken in time order autocorrelation of the errors will ordinarily be expected.

Generally an assumption of a first order autoregressive (AR-1) noise process with parameter $\phi$ $(-\infty < \phi < \infty)$ will be much closer to reality than an assumption of white noise. Analyzing in the Bayesian framework a particular data set (the CGK data) on the basis of a linear model with AR-1 noise (along the lines developed in Chapter 2), it was made very clear how very misleading inferences may be reached about the regression parameters $\theta_j$, $j = 1,2,\ldots,p$, if serial correlation is ignored (i.e. $\phi = 0$). Specifically it was seen, by looking at the joint posterior distributions of $\{\theta_j,\phi\}$ that the conditional distribution of $\theta_j$ changes not only in spread, but also in location as the conditional value for $\phi$ moves along its axis.

Realizing that dependence can create problems, one way in which data analysts traditionally have tried to get around this complication, is by differencing and then assuming, that the first differences have white noise errors, i.e. that the errors in the original observations follow a random walk.



Of course the AR-1 process includes a random walk as a special case, namely for $\phi = 1$. To assess the plausibility of that particular choice for $\phi$ the marginal posterior distribution of $\phi$, $p(\phi|\underset{\sim}{y})$, should be looked at. However the density at $\phi = 1$ must be determined separately since the model mean vanishes when $\phi = 1$, creating a singularity point. It was argued that $p(\phi|\underset{\sim}{y})$ does not in general go continuously through $\phi = 1$, and it was specifically proposed how to find the posterior density in that point.

The question of differencing was also approached from the point of view of seeing whether the evidence in the data supports spending one degree of freedom on differencing rather than on a fixed mean. In this treatment the ratio of the posterior probabilities for two reasonable alternative one parameter noise models was derived. One model assumes a fixed mean for an AR-1 noise, the other implies that the differenced series has AR-1 noise. In situations of doubt about which model should be the preferred one, marginal inference about the common linear parameters in the two models is actually still possible in the Bayesian framework without having to make a choice.

It would seem, that giving the data an opportunity to speak out in this quantitative way on whether a differencing is justified, would extend the appropriateness of a linear model analysis allowing for serial correlation to cover an even wider variety of real data sets.

Appendix, Other applications

In the analysis of the spirits data in Section 3.5.3 of Chapter 3 it might be wondered whether or not these data should be differenced, since the marginal posterior distribution of the autoregressive parameter $\phi$, labeled $p_n(\phi|\underset{\sim}{y})$ in Figure 3.6b, is located near $\phi = 1$. The density $\lim_{\phi \to 1} p_n(\phi|\underset{\sim}{y})$ is 8.92. Using (4.32) we find that the adjusted density at $\phi = 1$ is

$$p(\phi=1|\underset{\sim}{y}) = 8.92 \; e^{1/2} \left( \frac{.03298}{.03009} \right)^{-\frac{69-3-1}{2}} = .74 . \qquad (4A.1)$$

This much lower density discourage a differencing.

If the spirits data are analyzed according to model $H_1$, then the residual sum of squares is $SS_1 = .03284$, so that using (4.66) we find

$$\frac{P(H_0|\underset{\sim}{y})}{P(H_1|\underset{\sim}{y})} = \left( \frac{.02972}{.03284} \right)^{-\frac{69-3-1}{2}} = \frac{26.}{1.} , \qquad (4A.2)$$

which also testifies strongly against differencing.

An interesting application of the continuity adjustment derived in Section 4.2 presents itself in connection with the "artificial data" analyzed in Section 2.6.1 of Chapter 2. Figure 2.5b shows the marginal posterior distribution $p_n(\phi|\underset{\sim}{y})$; however since $\phi = .5$ is a singularity point, the density at that point must be found separately, namely from

$$p(\phi=\phi^*|\underset{\sim}{y}) = \lim_{\phi \to \phi^*} p(\phi|\underset{\sim}{y}) \; e^{1/2} \left( \frac{SS(\underset{\sim}{\theta}_{(1)}, \phi=\phi^*)}{SS(\underset{\sim}{\theta}, \phi=\phi^*)} \right)^{-\frac{n-p-1}{2}} \qquad (4A.3)$$

where $SS(\underset{\sim}{\theta}_{(1)}, \phi=\phi^*)$ and $SS(\underset{\sim}{\theta}, \phi=\phi^*) = \lim_{\phi \to \phi^*} SS(\underset{\sim}{\theta}, \phi)$ are computed as explained in Section 3.2 of Chapter 3.



For this example it is found that

$$p_n(\phi = .5|\underset{\sim}{y}) = 1.06 \; e^{1/2} \; .991 = 1.76 \qquad (4A.4)$$

so that the point $\phi = .5$, which is the true value for $\phi$, actually has the highest posterior density of any point.

## CHAPTER 5

### Summary

The analysis of linear models with independent, homoscedastic, normal noise (white noise) occupies a prominent position in statistics, and is used extensively as a tool in applied work in fields like business, economics and sociology as well as in the natural sciences and engineering. However when data have been collected sequentially in time or space the assumption of independence will ordinarily be unrealistic. If the assumptions are relaxed by introducing a single parameter to allow for serial correlation, then many real life data sets with autocorrelated errors may be approximately modelled.

Assuming that the observations (original or transformed) can adequately be represented by a linear model (with parameters $\underset{\sim}{\theta}$ ) whose noise term follows a first order autoregressive scheme (with parameter $\phi$), it was shown in Chapter 2 how in the Bayesian framework inferences can be drawn about $\underset{\sim}{\theta}$ and $\phi$ jointly, conditionally and marginally. Two AR-1 structures were considered; one (employing an additional starting parameter) covers explosive as well as stationary cases, the other assumes stationarity and reversibility. Approximately noninformative prior distributions for the parameters complementing the likelihood functions are suggested, in particular it is argued, that prior independence between $\underset{\sim}{\theta}$ and $\phi$ is not an appropriate assumption. In situations where the noise is clearly stationary, the inference about $\phi$ and in turn about $\underset{\sim}{\theta}$ appears very little affected by which of the two AR-1 schemes is employed. Using the more flexible not necessarily stationary model is generally recommended, although special care must be exercised in applying the locally approximately noninformative prior for $\phi$ .



As far as $\underset{\sim}{\theta}$ is concerned, it seems to make little difference in the analysis of real data however, if instead a convenient uniform prior for $\phi$ is adopted. The integration over $\phi$ leading to the marginal posterior distribution of $\underset{\sim}{\theta}$, must be carried out numerically, but excellent approximations may be obtained from a weighted sum of, say, five t-distributions.

Recognizing that the assumption of independence can be a crucial one, it has become a widespread practice in regression work, when observations are made in time order, to test for serial correlation using the well known Durbin Watson statistic. This test corresponds to determining whether a certain estimate $\check{\phi}$ of $\phi$ is significantly different from zero. Using a likelihood or Bayesian approach it is seen in Chapter 3 how better estimates for $\phi$ are obtained, and how approximate confidence or HPD intervals for $\phi$ are constructed. These approaches suggest testing procedures for the general hypothesis $\phi = \phi^{o}$; and in particular for $\phi^{o} = 0$ the alternatives may be viewed as competitors to the DW-test. It is demonstrated by simulation, that unless the linear model lacks a mean, the DW-test is comparable in power to the suggested alternatives. It is argued however, that it is really beside the point to test a null hypothesis of independence in such situations where serial correlation is to be expected. It was illustrated, that proceeding in the fashion implied by the DW testing approach carries a penalty, as namely the further analysis concerning $\underset{\sim}{\theta}$ may give very misleading results both if inferences about $\underset{\sim}{\theta}$ are made conditionally on $\phi = 0$ (when that hypothesis is "accepted") or if they are drawn conditionally on $\phi = \check{\phi}$ (if $\phi = 0$ is "rejected").

The Bayesian analysis of the regression model with AR-1 noise solves these problems. As far as $\underset{\sim}{\theta}$ is concerned, it is proposed that the full Bayesian analysis may be approximated by a conditional analysis using the maximum likelihood estimate $\hat{\phi}$ of $\phi$. ( $\hat{\phi}$ is found by estimating $\phi$ and $\underset{\sim}{\theta}$ simultaneously by least squares.) The resulting conditional inference is paralleled from a sampling theory point of view when inferences about $\underset{\sim}{\theta}$ are drawn as if $\phi = \hat{\phi}$.

The difficulties referred to above arise because of dependence of $\phi$ and $\underset{\sim}{\theta}$. The nature of this dependence is further clarified in Chapter 4, where for a particular example the joint posterior distributions of $\{\theta_j, \phi\}$ $j = 1,2,\ldots,p$ were considered. These plots demonstrate that the posterior distribution of $\theta_j$ conditional on $\phi = \phi^{\circ}$ not only changes in spread as $\phi^{\circ}$ moves along the $\phi$-axis, but also shifts location. One way in which data analysts traditionally have tried to get around the problems of serial correlation is by differencing and then assuming that the first differences have white noise errors, i.e. that the errors in the original observations follow a random walk. This is equivalent to assuming that $\phi = 1$ in the AR-1 noise model, and the plausibility of that particular value for $\phi$ may be assessed by studying the marginal posterior distribution of $\phi$. For models involving a mean this parameter vanishes for $\phi = 1$ creating a singularity point; and it is proposed how the density may be determined in such (distinct) points. The question of differencing was also approached in Chapter 4 from the point of view of determining whether the evidence in the data favors expending one degree of freedom on differencing or on a fixed mean. Specifically a quantitative rule was derived which expresses (in terms of relative posterior model probabilities) the evidence for and



against differencing prior to an analysis of the data along the lines of Chapter 2.

## REFERENCES


1. Abraham, B. and G. E. P. Box (1975), "Linear Models, Time Series and Outliers, 2: Outliers in Linear Models", Technical Report No. 437, Dept. of Statistics, Univ. of Wis., Madison, Wis.
2. Abrahamse, A. P. J. and A. S. Louter (1971), "On a New Test for Autocorrelation in Least Squares Regression", Biometrika, 58, 53-60.
3. Aitken, A. C. (1935), "On least squares and linear combinations of observations", Proc. Roy. Soc. Edinburgh, 55, 42-48.
4. Anderson, O. D. (1975), "A note on differencing autoregressive moving average (p,q) processes", Biometrika, 62, 521-523.
5. Anderson, R. L. (1954), "The Problem of Autocorrelation in Regression Analysis", JASA, 49, 113-129.
6. Anscombe, F. J. (1960), "Rejection of ontliers", Technometrics, 2, 123-146.
7. Anscombe, F. J. and Tukey, J. W. (1963), "The Examination and Analysis of Residuals", Technometrics, 5, 141-160.
8. Bartlett, M. S. (1947), "The Use of Transformations", Biometrics, 3, 39-52.
9. Berenblut, I. I. and G. I. Webb (1973), "A New Test for Auto-correlated Errors in the Linear Regression Model", JRSS series B, 35, 33-50.
10. Box, G. E. P. (1976), personal communication.
11. Box, G. E. P. and D. R. Cox (1964), "An Analysis of Transformations", JRSS Series B, 26, 211-252.
12. Box, G. E. P. and Henson, T. L. (1969), "Model fitting and discrimination", Tech. Report No. 211, Dept. of Stat., Univ. of Wis., Madison, Wis.

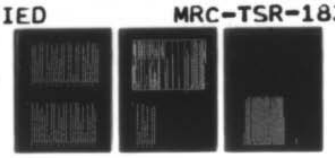




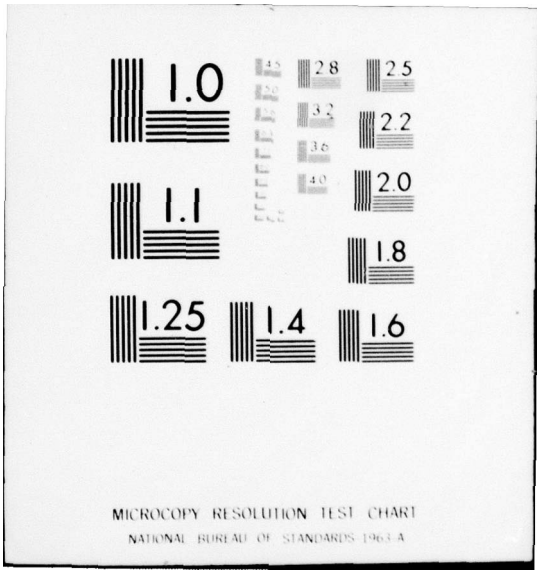
13. Box, G. E. P. and Hill, W. J. (1967), "Discrimination among mechanistic models", Technometrics, 9, 57-71.
14. Box, G. E. P. and G. M. Jenkins (1970), "Time Series Analysis, Forecasting and Control", Holden-Day.
15. Box, G. E. P. and P. Newbold (1971), "Some Comments on a Paper of Coen, Gome and Kendall", JRSS series A, 134, 229-240.
16. Box, G. E. P. and G. C. Tiao (1973), "Bayesian Inference in Statistical Analysis", Addison-Wesley Publishing Company.
17. Champernowne, D. G. (1948), "Sampling Theory Applied to Autoregressive Sequences", JRSS series B, 10, 204-242.
18. Cochrane, D. and G. H. Orcutt (1949), "Application of Least Squares Regression to Relationships Containing Auto-Correlated Error Terms", JASA, 44, 32-61.
19. Coen, P. J., E. D. Gomme and M. G. Kendall (1969), "Lagged Relationships in Economic Forecasting", JRSS series A, 132, 133-152.
20. Durbin, J. (1969), "Testing for serial correlation in regression analysis based on the periodogram of least squares residuals", Biometrika, 56, 1-15.
21. Durbin, J. (1970), "An Alternative to the Bounds Test for Testing for Serial Correlation in Least Squares Regression", Econometrica, 38, 422-429.
22. Durbin, J. (1970a), "Testing for Serial Correlation in Least-Squares Regression when some of the Regressors are Lagged Independent Variables", Econometrica, 38, 410-421.
23. Durbin, J. and G. S. Watson (1950), "Testing for Serial Correlation in Least Squares Regression. I.", Biometrika, 37, 409-428.

AD-A054 556 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
STUDIES IN THE ANALYSIS OF SERIALLY DEPENDENT DATA.(U)
MAR 78 L C PALLESEN DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1837 NL

2 OF 2
AD
A054556



END
DATE
FILMED
7-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

24. Durbin, J. and G. S. Watson (1951), "Testing for Serial Correlation in Least Squares Regression. II.", Biometrika, 38, 159-178.

25. Durbin, J. and G. S. Watson (1971), "Testing for Serial Correlation in Least Squares Regression, III.", Biometrika, 58, 1-19.

26. Durbin, J. (1960), "Estimation of Parameters in Time-Series Regression Models", JRSS series B, 22, 139-153.

27. Fraser, D. A. S. (1968), "The Structure of Inference", John Wiley and Sons, Inc.

28. Grary, R. C. (1970), "Relative Efficiency of Count of Sign Changes for Assessing Residual Autocorrelation in Least Squares Regression", Biometrika, 57, 123-127.

29. Hannan, E. J. (1957), "Testing for Serial Correlation in Least Squares Regression", Biometrika, 54, 57-66.

30. Haq, M. S. (1970), "Structural analysis for the first order auto-regressive stochastic process models", Ann. Math. Statist., 41, 970-978.

31. Haq, M. S. (1971), "Structural inference for the linear model with autoregressive error", J. Stat. Res., 5, 10-22.

32. Hildreth, C. (1969), "Asymptotic Distribution of Maximum Likelihood Estimators in a Linear Model with Autoregressive Disturbances", The Annals of Math. Stat., 40, 583-594.

33. Jeffreys, H. (1961), "Theory of Probability" 3rd ed., Oxford University Press.

34. Kanemasu, H. (1973), "Topics in Model Building", Ph. D. thesis, Dept. of Stat., Univ. of Wis., Madison, Wis.

35. Levenbach, H. (1972), "Estimation of autoregressive parameters from a marginal likelihood function", Biometrika, 59, 61-71.





36. Orcutt, G. H. and Cochrane, D. (1949), "A Sampling Study of the Merits of Autoregressive and Reduced Form Transformations in Regression Analysis", JASA, 44, 356-372.

37. Pierce, D. A. (1971), "Least squares estimation in the regression model with autoregressive-moving average errors", Biometrika, 58, 299-321.

38. Rao, C. R. (1965), "Linear Statistical Inference and its Applications", John Wiley and Sons, Inc.

39. Reiser, B. (1975), "Structural Inference for Linear Regression with Autocorrelated Errors", Stat. Hefte, 16, 85-104.

40. Schmidt, P. (1972), "A Generalization of the Durbin-Watson Test," Australian Economic Papers, 11, 203-209.

41. Smith, V. K. (1976), "The Estimated Power of Several Tests for Autocorrelation with Non-First-Order Alternative", JASA, 71, 879-883.

42. Sredni, J. (1970), "Problems of Design, Estimation, and Lack of Fit in Model Building", Ph.D. thesis, Dept. of Stat., Univ. of Wis., Madison, Wis.

43. Student (Gosset, W. S.) (1914), "The Elimination of Spurious Correlation Due to Position in Time or Space", Biometrika, 10, 179-181.

44. Theil, H. and A. L. Nagar (1961), "Testing the Independence of Regression Disturbances", JASA, 56, 793-806.

45. Tukey, J. W. (1957), "On the Comparative Anatomy of Transformations", Annals Math. Stat., 28, 602-632.

46. Tukey, J. W. (1960), "A survey of sampling from contaminated distributions", Contributions to Probability and Statistics: Volume Dedicated to Harold Hotelling. Stanford University Press.

47. Watson, G. S. (1967), "Linear Least Squares Regression", Ann. Math. Stat., 38, 1679-1699.

48 Whittle, P. (1953), "Estimation and Information in Stationary Time Series", Arkiv för Mathematik, 2, 423.

49. Wilks, S. S. (1962), "Mathematical Statistics", John Wiley and Sons, Inc.

50. Zellner, A. and G. C. Tiao (1964), "Bayesian Analysis of the Regression Model with Autocorrelated Errors", JASA, 59, 763-768.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER: 1837 ✓ | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle): STUDIES IN THE ANALYSIS OF SERIALLY DEPENDENT DATA | | 5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s): Lars Chr. Pallesen | | 8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024 ✓ |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: #4 - Probability, Statistics and Combinatorics |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709 | | 12. REPORT DATE: March 1978 |
| | | 13. NUMBER OF PAGES: 183 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report): UNCLASSIFIED |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Regression, Linear models, Bayesian approach, Serial correlation, Autoregressive noise, Nonstationary noise, differencing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The analysis of linear models with independent homescedastic, normal noise (white noise) occupies a prominent position in applied statistics. This report is concerned with the linear model analysis of data which cannot be assumed statistically independent because the data have been collected sequentially in time or space.

Assuming that the noise in a linear model (with p-dimensional parameter vector $\beta$) follows a first order autoregressive scheme (with parameter $\phi$) it is developed in Chapter 2 how in the Bayesian framework inferences can be drawn about

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Abstract (continued)

$\underset{\sim}{\theta}$ and $\phi$ jointly, conditionally and marginally. Two AR-1 schemes are considered, one covering explosive as well as stationary situations, the other assumes stationarity and reversibility a priori. The important task of choosing an appropriate joint prior distribution for the parameters is given special attention.

Recognizing that the assumption of independence can be a crucial one, it has become a widespread practice in much regression work, where observations are made in time order, to test for serial correlation using the well known Durbin-Watson test. This test corresponds to determining whether a certain estimate $\phi$ of $\gamma$ is significantly different from zero. It is shown in Chapter 3 that only in relation to a model lacking a mean do suggested alternative testing procedures show decisively greater empirical power than the DW-test. However it is argued that tests of a null hypothesis of independence should really not be carried out when serial correlation is to be expected. It is demonstrated that inferences about $\underset{\sim}{\theta}$ may be quite misleading if made conditionally on $\phi = 0$ (when that hypothesis is "accepted") or conditionally on $\hat{\phi}$ (when it is "rejected"). The Bayesian analysis does not suffer from these handicaps. The Bayesian inference about $\underset{\sim}{\theta}$ may be approximated by a conditional inference using the maximum likelihood estimate $\hat{\phi}$ of $\phi$ ( $\hat{\phi}$ is found by estimating $\phi$ and $\underset{\sim}{\theta}$ simultaneously by least squares), and this conditional analysis is paralleled in smapling theory framework when inferences about $\underset{\sim}{\theta}$ are drawn as if $\phi = \hat{\phi}$.

One way in which data analysts traditionally have tried to get around the problem of dependence, is by differencing and then assuming that the errors of the differences are independent. This is equivalent to assuming that $\phi = 1$ in the AR-1 noise model; and the plausibility of this particular value for $\phi$ may be assessed by studying the marginal posterior distribution of $\phi$. For models involving a mean this parameter vanishes for $\phi = 1$. This creates a singularity and it is shown in Chapter 4, how the density may be found at such (distinct) points. It is also developed in the Bayesian framework, how the appropriateness of differencing may be expressed as posterior model probabilities for two alternative noise models, one assuming the original observations to have AR-1 noise, the other that the differences have AR-1 noise.

LCP/db